Å skille klinten fra hveten

Hvordan kan ordbokaktuelle sammensetninger identifiseres?

Mikkel Ekeland Paulsen

Avhandling for graden philosophiae doctor (ph.d.)
Universitetet i Bergen
2024

UNIVERSITETET I BERGEN

Å skille klinten fra hveten

Hvordan kan ordbokaktuelle sammensetninger identifiseres?

Mikkel Ekeland Paulsen

Avhandling for graden philosophiae doctor (ph.d.)
ved Universitetet i Bergen

Disputasdato: 31.05.2024

År: 2024

Tittel: Å skille klinten fra hveten

Navn: Mikkel Ekeland Paulsen

Trykk: Skipnes Kommunikasjon / Universitetet i Bergen

Fagmiljø

Denne doktorgradsavhandlinga er skrevet ved Institutt for lingvistiske, litterære og estetiske studier (LLE), Det humanistiske fakultet, Universitetet i Bergen. Hovedveileder har vært professor Torodd Kinn ved samme institutt, og biveileder har vært Bård Uri Jensen, førsteamanuensis ved Institutt for nordisk språk og litteratur, Høgskolen i Innlandet. Jeg har vært tilknyttet Forskerskolen i lingvistikk og filologi ved Universitetet i Bergen, og jeg har deltatt på nasjonale og internasjonale konferanser underveis. I 2023 arrangerte den nevnte forskerskolen mesterklasse for meg med dosent Jan Svanlund fra Stockholms universitet. Våren 2022 gjestet jeg Institutionen för svenska, flerspråkighet och språkteknologi ved Göteborgs universitet med finansiering fra Vera och Greta Oldbergs stiftelse.

En fjerdedel av LLE sitt fireårige forskerutdanningsprogram består av pliktarbeid. Jeg har gjennomført pliktarbeid knyttet til faget nordisk og revisjonen av Bokmålsordboka og Nynorskordboka. Arbeidet har bestått av undervisning og veiledning av både bachelor-, master- og lektorstudenter i emner om metaforteori, leksikografi, semantikk, masterskriving og grunnleggende grammatikk. Deler av pliktarbeidet har også blitt brukt til formidling, blant annet på radioprogrammet NRK Språksnakk.

Forord

I dette prosjektet har jeg hatt et gunstig utgangspunkt både faglig og personlig. Jeg har dessuten nytt godt av de fremragende betingelsene som gjelder for doktorgradsstipendiater i Norge. I motsetning til mange andre steder har vi anstendig lønn, rikelig med driftsmidler, fleksible arbeidsforhold og faglig autoritet.

På den faglige siden er det spesielt veilederne mine, Torodd Kinn (hovedveileder) og Bård Uri Jensen (biveileder), som fortjener en stor takk. Det er gjennomgående deres skarpsynte og kunnskapsrike tilbakemeldinger som har forvandlet mine skisseaktige skriverier til publiserbare artikler. Uten dere ville det sannsynligvis båret galt avsted med både idiomatikk og statistikk.

Jeg vil videre takke Revisjonsprosjektet for å vise interesse for prosjektet og å gi meg stimulerende pliktarbeid, Clarino ved Paul Meurer for glimrende korpusressurser, Michael Eric Menk for tålmodig hjelp med å generere søkestatistikk og Emma Sköldberg for strålende vertskap ved Göteborgs universitet. Takk dessuten til Vera og Greta Oldbergs stiftelse for sponsing av forskeropphold i Göteborg. Og takk til Jan Svanlund for grundig lesing i forbindelse med mesterklassen.

Takk dessuten til Bill Evans, Gia Margaret, Maurice Ravel, Gjermund Larsen Trio og Selma French for musikk som formelig instituerer konsentrasjon og arbeidsdisiplin.

På den kollegiale siden fortjener Runa Falck, Anna Polster, kakelunsjgjengen på 108, adm.-kroken, Slabberaslaget og Stip-HF takk for kjærkomne sosiale atspredelser i løpet av arbeidsdagen.

På det personlige planet hører det med å takke Mamma og Pappa for uavlatelig støtte og oppmuntring gjennom oppvekst, skolegang, studier og doktorgradsarbeid, og brødra mine for ivrig radiolytting og språklig nysgjerrighet. I tillegg må jeg takke Kari for husly og hjemmekontor da taket rant i huet på oss.

Til slutt må jeg trekke fram Evelyn og Ola, som er utømmelige kilder til spas, og som har kullsviertro på meg og mine evner (i hvert fall hva gjelder det ervervsmessige). Selv etter den mest fenomenale arbeidsdag er det godt å komme hjem til dere.

Abstract

The aim of the present thesis has been to develop a rigorous and effective procedure for selecting compound-entries for general dictionaries. This aim builds on the premise that it is neither theoretically sound nor practically feasible to provide entries for every compound that has documented use in the Norwegian language. It is therefore necessary to find variables that enable one to extract a sample of compounds that is in line with linguistic, empirical and pedagogical considerations.

The thesis consists of three studies that in different ways fulfill the overall aim. These studies are joined by an overarching synopsis that summarises the project as a whole. Study 1 has a mainly qualitative design, whereas Studies 2 and 3 are based on quantitative analyses.

Study 1 investigates the variable semantic transparency of compounds. The compound literature includes many approaches to the definition and conceptualisation of how predictable or self-explanatory the meaning of a compound is given the meaning of its parts. In this study, five factors of semantic transparency, namely *extent of motivation*, *degree of motivation*, *internal disambiguation*, *schematic templates* and *schematic productivity*, are operationalised and combined in a model that is applied to estimate the degree of transparency of a selection of compounds. The model gives a fine-grained assessment of each compound's transparency and places them along a scale of eight tiers, where the compounds at the least transparent end have the strongest candidacy for being listed in dictionaries based on semantic transparency.

Among the five factors in the model, schematic productivity is particularly time-consuming to assess. Internal disambiguation is on the other hand relatively insignificant in that very few compounds are deemed more or less transparent based on this factor. The remaining factors do however appear to be both useful and effective tools for identifying the least transparent compounds among a group of candidates. There are however uncertainty tied to the operationalisation and aptness of these factors. One of the factors, degree of motivation, is evaluated in study 3.

The purpose of study 2 is to develop valid corpus methods for measuring frequency of use in the language as a whole. The study builds on a cross-validation analysis of

the performance of five different corpus measurements on 273 Norwegian compounds with varying types of corpus distributions. The analysis sheds light on the ability of each measurement to predict frequency of use in the language as a whole, and how the predictive accuracy of each measurement is affected by different types of corpus distributions.

The most important finding in study 2 is that the predictive accuracy of corpus frequency is dependent on the dispersion of a given n-gram. When measuring the corpus frequency of e.g. a compound, it is the dispersion of that compound that indicates the validity of that frequency measurement, that is whether the corpus measurement gives a true representation of the compound's frequency in the langugage as a whole. Among the dispersion measures in the study, *Deviation of Proportions* (DP) and *Juilland's D* show particularly promising results with respect to accuracy and stability.

Since these dispersion measures only indicate the proportionality with which a distribution is spread out across the corpus, they do not say anything about the magnitude of the distribution in question. For this reason, the most precise indications of the use of an n-gram in the language as a whole are made by dispersion and frequency estimates collectively. Frequency and dispersion estimates should therefore always be reported in order to mutually support one another. This ought to be a corpus convention analogous to the way a measure of variation, like standard deviation, is reported in statistic contexts to support estimates of central tendencies, like the mean.

Study 3 analyses *conditional inference trees* and *random forests* to identify the best linguistic and distributional predictors of look-up interest in the standard Norwegian dictionaries. To this end, I use a sample of approx. 1200 Norwegian compounds and an accumulated statistic of all effectuated query expressions from the same dictionaries in the time period 2016–2020.

The findings from Study 3 reveal a clear connection between corpus frequency and corpus dispersion on one hand, and look-up interest on the other. Notably, this relationship is unidirectional. High frequency and high dispersion are linked to high look-up interest, while low frequency and low dispersion do not correspond to decreased look-up interest. Hence, it is evident that diffusion in language use alone does not solely determine the variations in look-up interest.

Study 3 also tests the connections between various linguistic variables and look-up interest. A finding is that degree of motivation is somewhat associated with look-up interest, while part of speech affects the amount of variation that can be explained by the variables in the study. There is considerably more unexplained variation among binominal compounds than among non- and seminominal ones.

The studies collectively show that arriving at an ideal selection of compounds in dictiona-

ries partly involves identifying the most important linguistic and distributional variables, and partly involves finding valid ways to adapt these variables so that they can be easily applied in a lexicographic context. The search logs of the standard Norwegian dictionaries clearly demonstrate that user needs and interests in compound words are vast and varied, and that it will take many variables to explain the variation in user interest.

Furthermore, Study 3 shows that the traditional approach to selecting compounds, which presumably has relied on corpus frequency, semantic transparency and intuition, in the case of the Norwegian dictionaries has resulted in a relatively accurate selection of compounds with regard to search interest. To further improve accuracy, it is in all likelihood necessary to increase the resolution of the traditional variables. On the one hand, it is necessary to increase the resolution and validity of corpus investigations. At the very least, one must measure dispersion in addition to frequency. On the other hand, one must determine which variables constitute the seemingly accurate intuition that is currently applied.

When corpus frequency and semantic transparency have been used as more or less tacit variables for the selection of compounds, this presupposes that lexicographers ought to describe the vocabulary that is either conventional in usage or linguistically unconventional. In other words, compounds that either occur regularly in use or that deviate from the typical linguistic conventions or expectations in terms of meaning or structure.

Furthermore, it is obvious that within what is linguistically and empirically acceptable, one should try to capture most of the vocabulary that users are searching for. In this regard, Study 3 among others, shows that none of the variables examined in this thesis should be used to disqualify compounds from dictionary entries. Instead, one should operate with a set of qualifying factors that independently provide a basis for entry. In the final chapter of the thesis, the following set of qualifying factors is proposed, which attempts to capture words that are either conventionally used, linguistically unconventional or interesting to users:

1. Degree of diffusion
2. Degree of anomaly
3. Schematisation
4. Usualisation domain
5. Experential entrenchment
6. Attention value
7. Input to further word formation

To provide the most concrete fulfillment of the objectives of the thesis, the above variables are integrated into a lexicographic selection procedure. This procedure is also demonstrated on a selection of compounds with promising results.

Sammendrag

Målsetninga til denne avhandlinga har vært å utvikle en velfundert og effektiv prosedyre for seleksjon av sammensetninger til store allmennordbøker. Målsetninga utgår fra en kjensgjerning om at det verken er prinsipielt forsvarlig eller praktisk mulig for et ordbokverk å beskrive alle sammensetninger som er belagt i norsk språkbruk. Det er derfor nødvendig å finne variabler som hjelper en å skille ut det utvalget av sammensetninger som er mest hensiktsmessig fra et lingvistisk, empirisk og pedagogisk perspektiv.

Avhandlinga består av tre delstudier som på hver sin måte oppfyller deler av denne målsetninga. I tillegg kommer en kappe som med utgangspunkt i delstudiene og tidligere forskning svarer på den overordna problemstillinga. Delstudie 1 har hovedsakelig en kvalitativ innretning, mens delstudie 2 og 3 baserer seg på kvantitative analyser.

I delstudie 1 utforskes variabelen semantisk gjennomsiktighet. I litteraturen fins det mange innganger til å beskrive hvordan det varierer hvor forutsigbar eller selvforklarende betydninga til en sammensetning er, gitt delene den består av. I delstudien operasjonaliseres fem gjennomsiktighetsfaktorer: *Motivasjonsandel*, *motivasjonsgrad*, *intern disambiguering*, *skjematiske forbilder* og *skjematisk produktivitet*. De settes inn i en modell som samla anvendes til å beregne den overgripende gjennomsiktighetsgraden til et utvalg av sammensetninger med førsteleddene *svart-*, *tanke-* og *vandre-*. Modellen gir en finkorna inndeling av det aktuelle tilfanget av sammensetninger og fordeler dem langs en skala med åtte trinn, der sammensetningene i den minst gjennomsiktige enden formodentlig har det sterkeste ordbokkandidaturet basert på semantisk gjennomsiktighet.

Blant de fem faktorene i modellen i delstudie 1 utpeker skjematisk produktivitet seg som særlig tidkrevende å beregne for hver enkelt sammensetning. Intern disambiguering utpeker seg på sin side som en relativt uviktig faktor sammenlikna med de andre. De resterende faktorene virker på sin side som nyttige og effektive virkemidler for å skille ut de mest ugjennomsiktige sammensetningene i et utvalg. Det fins likevel spørsmål knytta til operasjonaliseringa og treffsikkerheten til disse faktorene. En av faktorene, motivasjonsgrad, blir imidlertid evaluert i delstudie 3.

Hensikten med delstudie 2 er å utvikle valide korpusmetoder for å måle frekvens i usus. Studien bygger på en kryssvalideringsanalyse av fem ulike korpusmål med et utvalg be-

stående av 273 norske sammensetninger med forskjelligarta korpusdistribusjoner. Analysen kaster lys over korpusmålenes evne til å predikere utbredelse i usus, og hvordan prediksjonsevnen blir påvirket av ulike typer korpusdistribusjoner.

Det viktigste funnet i delstudie 2 er at prediksjonsevnen til korpusfrekvens avhenger av korpusdistribusjonens spredningsgrad. Når man måler korpusfrekvensen til for eksempel en sammensetning, er det spredningsgraden til sammensetninga som indikerer validiteten til frekvensestimatet, altså om frekvensestimatet gir en presis framstilling av sammensetningas frekvens i usus. Blant spredningsmålene i studien ser *Deviation of Proportions* (DP) og *Juilland's D* ut til å gi de mest treffsikre og stabile prediksjonene av spredning i usus.

Spredningsmål sier imidlertid ingenting om størrelsesordenen til distribusjonen de anvendes på. Derfor er det en samla vurdering av korpusfrekvens og korpusspredning som best predikerer hyppigheten til et n-gram i usus. Analogt med den statistiske konvensjonen hvor man opplyser om standardavvik når man måler gjennomsnittet i et datautvalg, burde korpusspredning opplyses om i alle tilfeller der man rapporterer korpusfrekvens.

I delstudie 3 gjennomføres en inferenstre- og randomisert skoganalyse for å identifisere hvilke lingvistiske og distribusjonelle variabler som best predikerer søkeinteresse i standardordbøkene (Bokmålsordboka og Nynorskordboka). Til dette benyttes et utvalg på rundt 1200 sammensetninger og en akkumulert statistikk over benytta søkeuttrykk i søkefeltet til standardordbøkene i perioden 2016–2020.

Analysene i delstudie 3 indikerer at det er en positiv sammenheng mellom korpusfrekvens og korpusspredning på den ene siden og søkeinteresse på den andre. Korrelasjonen mellom disse variablene går imidlertid bare i én retning. Mens høy frekvens og jevn spredning er assosiert med høy søkeinteresse, er ikke lav frekvens og ujevn spredning nødvendigvis assosiert med lav søkeinteresse. Utbredelse i språkbruk er med andre ord ikke den eneste variabelen som forklarer variasjonen i søkeinteresse.

Delstudie 3 tester dessuten sammenhengen mellom en rekke lingvistiske variabler og søkeinteresse. Resultata fra dette indikerer at det er en ørliten positiv sammenheng mellom motivasjonsgrad og søkeinteresse. Dessuten forklarer variablene i studien vesentlig mer av variasjonen i søkeinteresse blant ikke- og seminominale sammensetninger enn blant binominale sammensetninger.

Delstudiene illustrerer samla at å komme fram til et hensiktsmessig sammensetningsutvalg i ordbøker dels dreier seg om å identifisere de viktigste lingvistiske og distribusjonelle variablene, og dels om å finne valide måter å tilpasse variablene på slik at de lett kan anvendes i en leksikografisk kontekst. Søkeloggene til standardordbøkene viser tydelig at brukerbehovet og -interessen for sammensatte ord er omfattende og variert, og at det trolig kreves mange variabler for å forklare variasjonen i brukerinteressen.

Delstudie 3 viser dessuten at den tradisjonelle inngangen til seleksjon av sammensetninger, som i stor utstrekning basererer seg på korpusfrekvens, semantisk gjennomsiktighet og intuisjon, i standardordbøkenes tilfelle har resultert i et relativt treffsikkert utvalg av sammensetninger med henblikk på søkeinteressen. For å kunne forbedre treffsikkerheten ytterligere dreier det seg trolig om at en må øke oppløsninga på de tradisjonelle variablene. På den ene siden må man øke oppløsninga og validiteten på korpusundersøkelsene; i det minste må man måle spredning i tillegg til frekvens. På den andre siden må man finne ut hvilke variabler som den noenlunde treffsikre intuisjonen er basert på.

Når korpusfrekvens og semantisk gjennomsiktighet har blitt brukt som mer eller mindre tause variabler for seleksjon av sammensetninger, vitner dette om et premiss om at leksikografer har ansvar for å beskrive det ordforrådet som enten er konvensjonelt i språkbruk eller ukonvensjonelt lingvistisk sett. Med andre ord sammensetninger som enten opptrer jevnlig i bruk, eller som enten betydningsmessig eller strukturelt avviker fra de typiske lingvistiske konvensjonene eller forventningene.

Dessuten er det innlysende at man innenfor det som er lingvistisk og empirisk akseptabelt, bør forsøke å fange mesteparten av det ordforrådet brukerne søker etter. I så henseende viser blant annet delstudie 3 at ingen av variablene som er satt under lupen i denne avhandlinga, bør benyttes til å ekskludere sammensetninger fra ordbokoppføring. Snarere bør man operere med et sett inklusjonskriterier, som uavhengig av hverandre gir grunnlag for oppføring. I avhandlingas nest siste kapittel foreslås det følgende settet med kvalifiserende egenskaper, som forsøker å fange ord som enten er bruksmessig konvensjonelle, lingvistisk ukonvensjonelle eller ettersøkte av brukerne:

1. Diffusjonsgrad
2. Anomaliseringsgrad
3. Skjematisering
4. Usualiseringsdomene
5. Erfaringsbasert innprenting
6. Oppmerksomhetsverdi
7. Innputt i videre orddanning

For å gi en mest mulig konkret oppnåelse av avhandlingas målsetning integreres de ovennevnte variablene inn i en operativ leksikografisk seleksjonsprosedyre. Denne prosedyren blir dertil demonstrert på et utvalg av sammensetninger, hvor resultatet framstår lovende.

Publikasjonsliste

Paulsen, Mikkel Ekeland. (2020). Svartsjuk tankelesing på vandresafari – en modell for bedømmelse av sammensatte ords gjennomsiktighet. *LexicoNordica, 27*, 161–187

Paulsen, Mikkel Ekeland. (2022). Assessing word commonness – Adding dispersion to frequency. *International Journal of Corpus Linguistics, 28*(3), 318–343. doi: 10.1075/ijcl.21037.eke

Paulsen, Mikkel Ekeland. (2023). Wheat or chaff? A compound selection model based on look-up data. *International Journal of Lexicography*, 306–324. doi: 10.1093/ijl/ecad013

Forkortelser

BOB – Bokmålsordboka
DP – Deviation of Proportions
Standardordbøkene – Samleterm for Bokmålsordboka og Nynorskordboka
LBK – Leksikografisk bokmålskorpus
NBbok – Nasjonalbibliotekets bokkorpus
NBavis – Nasjonalbibliotekets aviskorpus
NAOB – Det Norske Akademis ordbok
NOB – Nynorskordboka

Figurer

Tabeller

Innhold

Kapittel 1

Innledning

Den norske språket består både av urgamle, nedarva ordformer som er videreført og tilslipt gjennom århundrer, og av nyss ankomne ordformer som har blitt danna eller importert i nyere tid. Samla gir dette en forskjelligarta ordmasse hvor ikke alle ord er like etablerte, velkjente eller gjennomsiktige.

Det dynamiske ordtilfanget er en styrke for språkets brukere. De kan til enhver tid nyttiggjøre seg de ordformene som på en eller annen måte fins i språket, samtidig som de fritt og uten viderverdigheter kan importere eller innføre andre leksikalske uttrykk om de skulle behøve dem. Disse dynamiske sidene ved ordtilfanget påkaller spørsmål i leksikografien. Når kan leksikografene si at et ord ikke bare fra tid til annen anvendes i språket, men inngår i det? Hva ved ordas egenskaper eller anvendelse kan fortelle oss dette?

I denne avhandlinga vil jeg ta for meg en dynamisk og hyppig form for orddannelse i det norske språket, nemlig sammensetninger. Sammensetninger kjennetegnes ved at de, grovt sagt, består av to etablerte ordstammer (Johannessen, 2001). Den som kjenner ordstammenes betydning, har et godt utgangspunkt for å utlede sammensetningas betydning, selv ved første møte. Sammensetninger dannes derfor ofte spontant. Hver gang en språkbruker får behov for å betegne noe nytt, gi en ekstra spesifisering av noe eller komprimere et meningsinnhold, ligger sammensetningene latente. Dette fører til et språk som øyensynlig bugner over av sammensatte ord. For å skille mellom gull og gråstein må derfor leksikografene spørre seg: Hvor lenge har sammensetninga blitt anvendt i norsk? Er den kommet for å bli? Hvor anvendelig er den? I hvilken utstrekning blir den anvendt? Trenger brukerne å få den forklart, eller er den en selvfølgelig utbygging av sine deler?

1.1 Det sammensatte spørsmålet

Det er langt ifra trivielt å avgjøre hva som utgjør et gunstig sammensetningsutvalg i allmennordbøker, og videre hvordan man kan skjelne mellom gode og mindre gode ordbokord. Dette sammensatte spørsmålet avkrever svar på mer spesifikke spørsmål om hva som gjør et ordbokoppslag aktuelt, hvilke typer ord brukerne kan tenkes å ha vanskeligheter med eller være spesielt interesserte i, hvilke ord som inngår i det aktuelle språkets sentralordforråd, hvilke ord som betegner fenomener hvis ontologiske status fordrer ordbokoppføring, hvilke sammensatte betydninger som framstår som selvsagte gitt komponentene som utgjør dem, hvor mange ord man strengt tatt har plass eller kapasitet til å innlemme i et gitt ordbokverk, og hvem som primært er tenkt som målgruppe for en gitt ordbok. Med andre ord kan vi si at spørsmålet om hvordan og på hvilken bakgrunn man akkumulerer et hensiktsmessig utvalg med sammensetningslemma i allmennordbøker, er et sammensatt spørsmål. I denne avhandlinga vil jeg forsøke å gi det hittil grundigste svaret på dette spørsmålet.

Spørsmålet må besvares med en løsning som lett kan implementeres inn i allerede eksisterende leksikografiske praksiser og arbeidsmetoder. Selv om det naturligvis råder ulike praksiser mellom ulike leksikografiske prosjekt, kan vi tenke oss to ulike prinsipielle innganger til å sammenstille en endelig liste over sammensatte lemma. Én tilnærming ville vært å samle alle ordbokaktuelle sammensetningslemma i en liste som man ved ferdigstillelse av ordboka trimmer og korter ned til man er tilfreds med utvalgets innhold og størrelse. En slik metode er imidlertid noe upraktisk siden lista med potensielle sammensetninger ville vært milelang. En annen, og trolig mer praktisk, tilnærming er å vurdere sammensetninger i mindre grupper. For eksempel kan man i sammenheng med at man redigerer et grunnord, for eksempel *hest*, også vurdere sammensatte ord der *hest* inngår, for eksempel *hesterygg*, *hestehandel*, *arbeidshest* og *sengehest*.

Ved å bruke sistnevnte tilnærming vil det totale sammensetningsutvalget genereres gjennom tusentalls minianalyser. De ulike leksikografene i en redaksjon vil da gjøre daglige minianalyser som leder til at visse sammensetninger blir tatt med og visse utelatt med utgangspunkt i hva som virker hensiktsmessig innenfor rammene av en mindre ordmasse. For eksempel har jeg selv som ordbokredaktør gjort vurderinger av hvilke sammensetninger med forleddet *bade-* som fortjener plass i Bokmålsordboka (heretter *BOB*) (*Ordbøkene | Bokmålsordboka og Nynorskordboka*, 2023). Disse vurderingene ble hovedsakelig gjort på bakgrunn av interne forhold og utbredelsen til ulike *bade*-ord, ikke på bakgrunn av et sett med generelle kriterier for sammensetninger i BOB. Likevel kommer også mer allmenngyldige lingvistiske vurderinger i spill innenfor konteksten av de ulike badeorda. For eksempel kan man vurdere det slik at sammensetningene *badeand* og *badedrakt* bør få oppføring ettersom de betegner noe annet enn henholdsvis en typisk and

eller en typisk drakt. I dette tilfellet vil det være selvstendige lingvistiske egenskaper, en slags semantisk anomali, som motiverer oppføringa til enkelte badeord. Dermed ser vi at både generelle regler for sammensetninger og relative forhold mellom ørsmå grupper av sammensetninger kan spille inn på det endelige sammensetningsutvalget til en ordbok.

Det overordna spørsmålet i denne avhandlinga er derfor både et spørsmål om helt konkrete og dagligdagse leksikografiske minivurderinger og et spørsmål om større og mer overordna prinsipielle, teoretiske og empiriske avveininger.

1.2 Hvorfor spørsmålet?

I dette delkapittelet vil jeg komme inn på ulike sider ved sammensetting og sammensetninger i norsk som motiverer det overordna spørsmålet for avhandlinga. I første omgang vil dette dreie seg om sammensetningene som en bestanddel av det objektet allmennordbøker tar sikte på å beskrive, altså en språklig varietet. Hva slags avtrykk har sammensettingsfunksjonen i norsk, og hvordan skal eller kan ordbøker gjenspeile dette?

Alle ordbøker har et spesifikt språklig objekt de har som formål å beskrive. I mange tilfeller er det ordtilfanget innenfor en gitt språklig varietet med en viss temporal avgrensning som utgjør dette objektet (Fjeld & Vikør, 2008, 136). Hvis vi legger dette til grunn, kan vi umiddelbart slutte at det fins et viktig misforhold mellom ordboka som objekt og den språklige varieteten som objekt, nemlig at ordboka er en finitt og tydelig avgrensa størrelse, mens ordtilfanget til en språklig varietet er prinsipielt uavgrensa, i alle fall hvis det har produktive orddanningsmekanismer som avleding og sammensetting. Dette misforholdet «tvinger» leksikografer til å beskrive noe som kan tjene som en representasjon av den språklige varieteten. Dette kan for eksempel være noe som på et eller annet nivå kvalifiserer som «de vanligste» orda og ordelementa i varieteten, ord som inngår i et slags sentralordforråd, eller de hyppigst brukte orda i den aktuelle varieteten. Leksikografene må i alle fall nøye seg med å beskrive et utvalg av vokabularet til den språklige varieteten, og i særdeleshet må de gjøre et utvalg innenfor massen av sammensatte ordformer.

Det er flere grunner til at det må gjøres et særlig utplukk blant sammensetninger. Siden sammensetninger til syvende og sist består av minst to simplekse og udelelige ordstammer (se videre forklaring i delkapittel 2.3), gir det seg selv at ordstammene oppfattes som mer grunnleggende og sentrale byggesteiner i språket enn sammensetningene de inngår i, på samme måte som poteter og melk er mer grunnleggende enn potetmos. I tillegg er det også en kjensgjerning at simplekser som uttrykk har en mer arbitrær forbindelse til sin betydning enn en sammensetning, som ofte er motivert av sine komponenters allerede etablerte betydning. Det er helt umulig å utlede betydninga til uttrykket *flagg*

fra uttrykkets form aleine, mens betydninga til uttrykket *flaggstang* kan utledes dersom en kjenner betydninga til komponentene det består av. Derfor gir det god mening at de fleste ordbøker er mer inkluderende overfor simplekser enn komplekser.

Et annet poeng, som er til dels overlappende med det ovenfor, er at sambandet mellom et simpleks og en gitt betydning nødvendigvis må være konvensjonalisert for at det aktuelle simplekset skal kunne fungere optimalt i verbal kommunikasjon. Med andre ord må betydninga til uttrykket være etablert innenfor det språkmiljøet hvor uttrykket ytres. For en sammensetning holder det at betydninga til de simplekse eller komplekse leddene den består av, er konvensjonaliserte. Om man for eksempel utbryter at «nå hadde det gjort seg med en snakkepause», behøver ikke mottakeren av dette budskapet å ha vært tidligere eksponert for sammensetninga *snakkepause* for å kunne utlede at dette i konteksten betyr noe sånt som 'pause fra/til snakking'. Sammensetninga trenger med andre ord ikke være konvensjonalisert for å bli forstått, og det er derfor større grunn til å inkludere belagte simplekser enn belagte sammensetninger. Et videre poeng her er også at det ville være direkte misvisende for en ordbok å gi oppslag til ikke-konvensjonelle sammensetninger, da et ordbokoppslag i seg selv indikerer at sambandet mellom oppslagsordet og dets betydningsangivelse er konvensjonell. Det er naturligvis ikke en ordboks mandat å beskrive det totale hypotetiske ordtilfanget som fins innenfor den språklige varietetens system, men heller å beskrive det ordtilfanget som på et eller annet nivå er aktualisert (Atkins & Rundell, 2008; Fjeld & Vikør, 2008, 156).

En videre grunn til at ordbøker med fordel lar mange sammensetninger ligge i skuffen, er at det er praktisk umulig (og prinsipielt uforsvarlig) å inkludere alle. I et språk som norsk, som ifølge Eik (2019, 19) har et typisk germansk produktivt sammensetningssystem, er omfanget av ulike sammensatte ordformer i prinsippet uavgrensa og i praksis uoverskuelig digert. Et illustrerende eksempel finnes hos De Smedt (2021), som slår fast at det over en periode på 139 dager i 2021 ble produsert over 1200 ulike nye sammensetningstyper i norske nettaviser med forleddet *corona/korona*. På det meste ble over 200 ulike typer anvendt samme dag. Den høye produktiviteten til sammensetting som orddanningsfunksjon lar seg også lett spore i trunkerte korpussøk. Om man for eksempel søker på «arbeid*» i Leksikografisk bokmålskorpus (Fjeld, Nøklestad & Hagen, 2020), vil man finne omtrent 2700 unike leksem, hvorav de aller fleste er sammensetninger. Videre oppgir Kjelsvik (2017) at over 60 % av oppføringene i Norsk ordbank for bokmål og nynorsk er sammensetninger.[1]

Tradisjonelt har ordbøker vært begrensa av størrelsen på og sideantallet til den fysiske boka. Ordmengden må her tilpasses etter hvor stor og tung boka kan være før det

[1] Ordbankene er leksikalske databaser som inneholder grunnord og bøyningsmønstre for henholdsvis bokmål og nynorsk. Se mer informasjon på nettsiden www.uib.no/ub/spesialsamlingene/160646/ordbøker.

blir ukurant å bruke den, og hvor liten og tettpakka skrifta kan være før det blir for anstrengende å lese den. I digitale ordbøker er naturligvis ikke denne faktoren like begrensende, men det fins like fullt andre begrensninger knytta til hvor mye tid, penger og arbeidskraft man har tilgjengelig til både kompilering og vedlikehold av artiklene i en ordbok.

I tillegg til de elementære kjensgjerningene at sammensetningsmekanismen er produktiv i norsk, og at leksikografer som beskriver norsk derfor må gjøre et utvalg av sammensetninger, tilkommer følgende kjensgjerninger som vil utdypes nærmere i kapittel 2.

- Sammensetninger har ulik etableringsgrad (se inngående diskusjon i delkapittel 2.4).
- Sammensetninger har ulik gjennomsiktighetsgrad (se inngående diskusjon i delkapittel 2.5).
- Sammensetninger har ulik bruksfrekvens (se inngående diskusjon i delkapittel 2.6).

1.3 Problemstilling og forskningsspørsmål

Med situasjonen som beskrives ovenfor, i mente formuleres følgende problemstilling med tilhørende forskningsspørsmål for denne avhandlinga:

- Hva utgjør et gunstig sammensetningsutvalg i allmennordbøker, og hvilke metoder og variabler er hensiktsmessige for å identifisere medlemmer av dette utvalget?

Problemstillinga besvares av tre delstudier og denne sammenfattende teksten (heretter kalt *kappa*). Den er videre utdypet i en rekke forskningsspørsmål som her er fordelt etter delstudie:

- **Delstudie 1**: **Svartsjuk tankelesing *på* vandresafari – *en modell for bedømmelse av sammensatte ords gjennomsiktighet***
  - Hvordan kan en operasjonalisere semantisk anomaliseringsgrad som leksikografisk utvelgelseskriterium for sammensatte ord?
- **Delstudie 2: *Assessing word commonness – Adding dispersion to frequency***
  - Hvordan kan en måle utbredelse i usus?

  - Hvordan kan en anvende korpusundersøkelser til å identifisere sammensetninger med høy etableringsgrad?

- **Delstudie 3: *Wheat or Chaff? A Compound Selection Model Based on Look-Up Data***

  - Hvilke kvantitative og kvalitative variabler predikerer brukernes søkeinteresse?
  - Hvilken kombinasjon av variabler kan brukes til å filtrere ut det mest treffsikre utvalget av sammensetninger med hensyn til søkeinteresse?
  - Hvor godt modellerer en slik kombinasjon av variabler søkeinteressen sammenlikna med den gjeldende lemmalista av sammensetninger i standardordbøkene?

- **Kappa**

  - Hvilke variabler spiller inn på sammensetningers ordbokrelevans?
  - Hvordan kan disse variablene anvendes for å skille ut et hensiktsmessig sammensetningsutvalg?

For å besvare de ovenstående forskningsspørsmåla anvendes gjennomgående en bruksbasert tilnærming. Det teoretiske rammeverket er ellers intensjonelt eklektisk, men har et betydelig innslag av kognitive innfallsvinkler. Datagrunnlaget beskrives kort i det følgende.

1.4 Datagrunnlag

Det empiriske datagrunnlaget for denne avhandlinga består i hovedsak av det balanserte korpuset Leksikografisk bokmålskorpus (LBK) (Fjeld et al., 2020) og innhold og søkestatistikk fra de offisielle ordbøkene for bokmål og nynorsk, Bokmålsordboka (BOB) og Nynorskordboka (NOB) (Ordbøkene, 2023). Samlebetegnelsen for ordbøkene er *standardordbøkene*.

Standardordbøkene ble først utgitt i 1986 og har seinere utkommet i flere trykte utgaver (Ordbøkene, 2023). Ordbøkene fins i dag i oppdatert versjon som en digital ordbok der både BOB og NOB er tilgjengelige fra samme søkegrensesnitt. Ordbøkene har fra 2018 vært gjenstand for en systematisk revisjon, der endringer har blitt fortløpende publisert på ordbokene.no (Revisjonsprosjektet 2023). Per 25 oktober 2023 består BOB og NOB av henholdsvis ca. 81 000 og ca. 96 000 oppslagsord. Ordbøkene har som formål

å dekke offisiell gjeldende rettskrivning med en lemmaliste som representerer det sentrale ordforrådet i disse skriftspråka de siste 50 årene (Ordbøkene, 2023).

Et korpus er dypest sett en samling språkobservasjoner. På lik linje med at en klimaforsker kan indusere globale klimaforandringer ut fra observasjoner av istykkelse og havtemperaturer, eller en medisinsk forsker kan indusere virkninger av et legemiddel ut fra observasjoner av symptomer hos en gruppe testpersoner, så kan lingvister indusere mønstre i språkbruk og regler i et språksystem ut fra en nøye utvalgt samling av autentiske forekomster av språket i bruk. Denne samlinga, altså korpuset, er på et eller annet nivå en representasjon av en språklig varietet eller en distinkt kategori innenfor denne. Om målobjektet, altså det fenomenet korpuset skal representere, er finitt, vil det være mulig å kompilere et fullstendig korpus.[2] De fleste korpus må imidlertid anses for å være statistiske utvalg som representerer en større, ofte uavgrensa, størrelse, som for eksempel et språk. Hva slags hypoteser man kan teste ved hjelp av et korpus, er betinga av nettopp forholdet mellom korpuset og dets målobjekt.

Leksikografisk bokmålskorpus er et innholdsmessig balansert korpus som består av omlag 100 millioner ord. Det inneholder bokmålstekster fra perioden 1985 til 2013 som er annotert med både grammatisk og opphavsmessig informasjon.

Selv om avhandlinga med dette datagrunnlaget knytter seg tett til ordbokverka BOB og NOB, og til korpuset LBK, vil funna ha relevans for andre allmennleksikografiske prosjekt innenfor språk med produktive orddanningsfunksjoner som sammensetting eller avleding.

1.5 Struktur

Kappa er bygget opp på følgende måte: Det inneværende kapittelet gir en innledning til og begrunner temaet for avhandlinga og den leksikografiske problemstillinga den springer ut av. Videre presenteres overordna problemstilling for avhandlinga sammen med en rekke forskningsspørsmål som er sortert etter hvilken delstudie som er ment å besvare dem.

Kapittel 2 gir både en forskningshistorisk oversikt og diskuterer de språkteoretiske og metodiske forutsetningene som avhandlinga bygger på. Det legges særlig vekt på ulike innganger til å beskrive sammensetningers morfologi, semantikk og varierende etableringsgrad, og på epistemiske spørsmål knyttet til korpus og korpusmetodikk.

I kapittel 3 redegjør jeg for den konkrete metodiske innretninga, samt datagrunnlaget og -utvalget til delstudiene i avhandlinga.

[2]For eksempel er det både prinsipielt og praktisk mulig å innlemme hele forfatterskapet til Sigrid Undset i et korpus som representerer nettopp dette.

I kapittel 4 sammenfattes bakgrunn og funn for hver av delstudiene.

I kapittel 5 samles først trådene fra delstudiene og kappa, før den overordna problemstillinga for avhandlinga og forskningsspørsmåla knytta til kappa blir besvart.

I kapittel 6 avsluttes avhandlinga med drøfting av kunnskapsstatus, forslag til videre forskning og en kort oppsummering.

De publiserte artiklene som utgjør delstudie 1–3, ligger vedlagt etter kappa.

Kapittel 2

Teoretisk bakgrunn

2.1 Leksikografi

Fjeld og Vikør (2008, 18) skiller mellom leksikologi, som innbefatter teoretisk og empirisk undersøkelse av ord og ordforråd, og leksikografi, som er den praktiske beskrivelsen av ordforråd i ordsamlinger. Selv om leksikografer hovedsakelig er forbundet med den praktiske beskrivelsen, må de uvegerlig bedrive leksikologi i tillegg, for forut for enhver leksikografisk beskrivelse kommer nødvendigvis en leksikologisk undersøkelse. Denne avhandlinga har et delt leksikologisk og leksikografisk utgangspunkt. På den ene sida er det en målsetning å utvikle arbeidsmetoder innenfor praktisk leksikografi, på den andre sida må det være leksikologiske innsikter som ligger til grunn for disse arbeidsmetodene.

Leksikografiske undersøkelser innebærer som regel *metodetriangulering*. I leksikografisk sammenheng består trianguleringa av at både kvantitative og kvalitative tilnærminger legges til grunn for å fatte beslutninger. I delkapittel 2.4 til 2.6 diskuteres de mest sentrale variablene som inngår i denne metodetrianguleringa. Et mål med variabelutvalget må være at det resulterer i et hensiktsmessig lemmautvalg.

Leksikologiske undersøkelser behøver videre en teoretisk forankring. Om undersøkelsene er ment å tjene som grunnlag for allmennleksikografi, er det vesentlig at den teoretiske forankringa er hensiktsmessig med tanke på dette. To viktige faktorer i denne sammenhengen er 1) at den jevne språkbruker ikke har oversikt over ulike lingvistiske skoler, og 2) at allmennordbøker ofte oppfattes som normgivende om ordtilfang og ordas betydninger (Atkins & Rundell, 2008, 2). Disse kjensgjerningene fordrer at leksikografer som redigerer ordbøker på vegne av og til nytte for fellesskapet, bør anta en mest mulig teorinøytral (eller teoriovergripende) og konsensusprega tilnærming til å beskrive språket.

2.1.1 Allmennordbokas raison d'être

Leksikografer har på et eller annet nivå et terreng å kartlegge. For at kartet skal gi en tilfredsstillende framstilling av terrenget, må nødvendigvis de største veiene, fjella, skogene, vassdraga og bebyggelsene tegnes inn. Nettopp dette – å gi en tilfredsstillende framstilling av språket og dets inventar – er en av allmennordbokas hovedoppgaver. Før man kan begynne å diskutere hvordan man skal få til dette, må man ha en klar idé om hva det angjeldende språklige objektet består av. For en synkron allmennordbok er det for det første nødvendig å avgrense språket temporalt (Fjeld & Vikør, 2008, 159). Siden en allmennordbok forsøker å gi informasjon om det ordtilfanget som eksisterer i det nåtidige språket, må definisjonen av *nåtid* nødvendigvis ha en bakre grense, et svar på når nåtiden begynte. Det ligger videre i ordet *allmennordbok* at det språklige objektet først og fremst er det som påtreffes i allmenne fora, altså språklige former som på et eller annet vis er å anse som felleseie i språksamfunnet. Likevel er det ikke noe vannskille mellom verken nåtid og fortid, eller mellom allmennspråk og spesialspråk. At en ordbok tar mål av seg til å være allmennspråklig og synkron, er mer å betrakte som en primær ambisjon og målsetning for det konkrete ordbokprosjektet enn et absolutt dogme som ordboka etterlever til punkt og prikke.

Videre må det tas stilling til om allmennordboka skal ha et deskriptivt eller normativt siktemål. I praksis er de fleste allmennordbøker både deskriptive og normative samtidig (se for eksempel Bergenholtz og Bøgelund (2002); Fjeld (2002); Fjeld og Vikør (2008)). Som Bergenholtz og Bøgelund (2002) påpeker, fins det dessuten forskjeller i om en ordbok er eksplisitt eller implisitt normativ eller deskriptiv. Det er ikke alltid en ordbok forklarer eller spesifiserer hvilke opplysninger som er ment å være for eksempel normative. Og uansett hvor eksplisitt deskriptiv den er, kan den komme til å bli oppfatta som normerende på visse punkt. Som Fjeld (2002) påpeker, er ordbøker med på å definere hva som befinner seg i språkets innmark, hvilke ord og uttrykk som er lagt under plogen. Brukere av språket kan derfor oppfatte ordbokførte ord som mer normerte og konvensjonaliserte enn ikke-ordbokførte ord.

I norsk sammenheng fins det en relativt stor språklig utmark. Standardordbøkenes opplysninger om staving og bøyning er normative i den forstand at de aktuelle rettskrivningsopplysningene foreskrives av Språkrådet. Dessuten formidler inntak av et ord i standardordbøkene at det er fullt brukbart innenfor normert bruk av bokmål og nynorsk. Men selv om alle oppførte leksem er normerte i de respektive skriftspråka, betyr ikke det at alle ikke-førte leksem er unormerte. Regelmessige sammensetninger og avledninger er for eksempel normerte så lenge deres bestanddeler er det. Brukere kan likevel tolke dem som del av den språklige utmarken, som ellers består av blant annet unormerte skrivemåter og importord, samt unormerte dialektale, sosiolektale eller etnolektale

former. En av hovedoppgavene til en allmennordbok er dermed å avgrense, definere og beskrive det som av brukerne blir oppfatta som den språklige innmarken.

En annen viktig oppgave for en allmennordbok er simpelthen å tjene som et nyttig verktøy for sine brukere. Den vanligste bruken av ordbøker baserer seg ikke på at de er viktige kulturhistoriske og dokumentariske artefakter, men at de er verktøy i forbindelse med språklig produksjon og resepsjon. Ordboka skal kunne gi viktige opplysninger om typisk bruk, betydning, opprinnelse, staving og bøyning til alt fra profesjonelle skribenter til innlærere av språket, både i produktive og reseptive kontekster. Studier av ordbokbruk viser at brukere benytter ordbøker til å finne ulike typer opplysninger, og en og samme bruker kan gjerne behøve å slå opp den normerte bøyninga av et ord og den typiske bruken av et annet (Pilke, 2008). Siden ordboka for mange brukere er premissleverandør for hva det går an å skrive eller si på et språk, bør allmennordbøker forsøke å fange mest mulig av det som rettmessig tilhører den språklige innmarken, selv om de i tilfellet med blant annet sammensetninger neppe kan fange hele.

En måte å ivareta ordbokas verktøyfunksjon på er å være lydhør overfor brukerne og å studere deres atferd i ordbøker. Når det gjelder spørsmålet om ordtilfanget i ordboka, er det aktuelt å studere hvilke ord brukerne slår opp. Mange ord bør så klart være med uavhengig hvor ofte brukerne søker på dem. Dette er ord som uomtvistelig inngår i den språklige varieteten ordboka beskriver, eller ord som trengs til å løse interne leksikografiske oppgaver som å henvise mellom artikler eller inngå i definisjonsformuleringer av andre ord. Utover disse uomtvistelige ordbokinnslaga kan søkestatistikk gi verdifull informasjon om hvilke ord ordboka bør ta med nettopp for å tjene som en best mulig ressurs for brukeren.

2.2 Bruksbasert teori

Undersøkelsene i denne avhandlinga har prinsipielt et eklektisk teoretisk utgangspunkt ved for eksempel at både generativ og kognitiv lingvistisk forskning anvendes. Likevel er tilnærminga gjennomgående bruksbasert. Ei slik tilnærming innebærer at språkbruk er den viktigste kilden til kunnskap om språket på alle nivåer, det være seg fonologi, morfologi, syntaks og semantikk med flere (Evans & Green, 2006, 108). Moderne leksikografi er også grunnleggende bruksbasert ved at det er bruken, *usus*, som i hovedsak avgjør hvilke morfologiske og semantiske opplysninger en ordbok gir (Fjeld & Vikør, 2008, 156). I visse tilfeller kan selvsagt språksystemiske hensyn ha forrang foran empiriske funn, men i all hovedsak er det empiri fra autentisk språkbruk som informerer leksikografiske vurderinger. Et sentralt postulat innenfor bruksbasert teori er at grammatikken på individnivå formes av bruk (se f.eks. Fløgstad (2022); Tomasello (2006)). Leksikografi dreier

seg imidlertid ikke om å beskrive individuelle grammatikker, men snarere ordforråd og grammatikk på språksamfunnsnivå, som også er forma av bruk.

I denne avhandlinga vil jeg ta utgangspunkt i at det fins et nært forhold mellom individ og samfunn hva gjelder grammatikk og ordforråd. Siden språkkompetanse på individnivå forutsetter evnen til å kunne kommunisere vellykka med andre individer av samme språksamfunn, må det nødvendigvis være mye overlapp mellom ord- og konstruksjonsinventar og grammatiske regler på tvers av individer. Schmid (2020, 1) påpeker at vellykka kommunikasjon forutsetter at språkbrukere på et eller annet nivå følger de samme konvensjonene, noe som igjen forutsetter at de har kjennskap til disse. En kjensgjerning er likevel at ordtilfanget og grammatikken til enhver språklig varietet med god margin overskrider ordforrådet og grammatikken til selv den mest elokvente språkbruker. Det ville for eksempel vært sensasjonelt om et individ var i stand til å liste opp og bøye – eller simpelthen forstå – samtlige 81 000 oppslagsord i BOB (per 25.10.2023). Derfor kan ikke leksikografer ta utgangspunkt i bare sin egen lingvistiske kompetanse – de må undersøke hva slags forråd og grammatikk som avtegner seg i kommunikasjon mellom språkbrukerne, det vil si i språkbruken.

Det bruksbaserte perspektivet viser seg altså i det objektet leksikografien søker å beskrive. Som Fjeld og Vikør (2008, 156) poengterer, er det leksikografers oppgave å beskrive det aktualiserte ordtilfanget i en varietet, ikke å beskrive et hypotetisk, systemgenerert ordforråd. Spørsmålet om hvilke ord en språklig varietet innehar, besvares ved å studere hvilke ord brukere av varieteten bruker. Videre vil det bruksbaserte perspektivet i denne avhandlinga også omfatte språkbrukere som ordbokbrukere. Parallelt med at språkbruk vil bli undersøkt ved hjelp av korpus (se underkapittel 3.1.2), vil ordbokbruk bli undersøkt ved hjelp av søkestatistikk (se underkapittel 3.2.3). Et siste bruksperspektiv ligger på leksikografiske vurderinger. Selv om leksikografer ikke sånn sett *bruker* ordboka og innholdet i den, er det likevel slik at ordbokas artikkelliste med innhold gjenspeiler resultatet av leksikografiske aktiviteter og vurderinger, på lik linje med at korpus gjenspeiler resultatet av språklige aktiviteter og søkestatistikk gjenspeiler ordbokbrukernes aktiviteter. Det er likevel en gradsforskjell i spontanitet mellom konvensjonelle språklige ytringer og søk i ordbøker på den ene siden og ordbokredigeringer på den andre. Selv om spontan språkbruk intuitivt kan oppleves som mer autentisk enn planlagt, korrekturlest og gjennomarbeida språkbruk, gir det liten mening å si at spontane leksikografiske foreteelser er mer autentiske enn mer planlagte foreteelser. Det er rettere sagt slik at leksikografi som felt er kjennetegna av nøye uttenkte og planlagte «ytringer» (i form av artikler) som blir til gjennom redaksjonelt samarbeid.

Ifølge Schmid (2020, 13) utgjør språkbruk også et arnested for prosessene han kaller *konvensjonalisering* og *innprenting*. Termene denoterer henholdsvis de sosiale og kognitive prosessene som utgjør «the linguistic system». Poenget er at språkbruken både påvir-

ker og blir påvirket av disse prosessene. Konvensjonelle og bredt innprenta enheter har større sjanse for å bli tatt i bruk og såleis videre konvensjonalisert og innprenta. Derfor har innprentings- og konvensjonaliseringsmodellen til Schmid (2015, 2020) en tilbakekoplingsfunksjon som gjør at utputten, altså resultatet av gjentatt språkbruk, informerer innputten, altså den til enhver tid gjeldende sannsynligheten for at noe blir tatt i bruk på tvers av kontekster. I det følgende beskrives begrepa konvensjonalisering og innprenting nærmere.

2.2.1 Konvensjonalisering

Konvensjonalisering kan forstås som en prosess der språklige uttrykksmåter blir gradvis innarbeida i en språklig varietet gjennom kollektive mellommenneskelige, ofte implisitte, overenskomster. Med andre ord dreier konvensjonalisering seg om «hur en viss språkgemenskap koordinerar sig om ett visst uttryckssätt och dess betydelse» (Svanlund, 2009, 38). I Saussures logikk er konvensjoner det som kompenserer for formens arbitrære form vis-a-vis innholdet den betegner (Ferdinand de Saussure, i Schmid (2020)). Innholdet i et uttrykk kan ikke avleses i uttrykkets form, derfor må assosiasjonen mellom form og innhold forhandles fram, konvensjonaliseres og læres:

> a *convention* [is defined] as a mutually known regularity of behaviour which the members of a community conform to because they mutually expect each other to conform to it. (Schmid, 2020, 88)

Konvensjoner eksisterer som taus og delt kunnskap mellom medlemmer av et språksamfunn, dvs. brukere av en språklig varietet. Sambandet mellom et uttrykk og et begrep er konvensjonalisert hvis for eksempel språkbrukere med en viss suksess kan framkalle begrepet mentalt hos andre gjennom å bruke uttrykket (semasiologisk konvensjonalisering), eller hvis man for eksempel gjennom miming eller forklaring av begrepet kan få andre til å bruke uttrykket (onomasiologisk konvensjonalisering). Konvensjonalitet er uansett et gradsfenomen siden et uttrykk til enhver tid kan bli ytterligere konvensjonalisert gjennom at for eksempel konkurrerende uttrykk taper terreng, eller også mindre konvensjonalisert gjennom at for eksempel konkurrerende uttrykk vinner terreng.

Konvensjonalisering kan videre brytes opp i de mer eller mindre distinkte prosessene *usualisering* (usualization) og *diffusjon* (diffusion) (Schmid, 2020, 87). Førstnevnte refererer til prosessen der konvensjoner etableres, sistnevnte til spredninga av denne konvensjonen. Forholdet mellom fagtermer og allmennord kjennetegnes for eksempel av ulik diffusjon, men samme usualisering. Termene *leksem* og *ord* (mine eksempler) har begge en konvensjonalisert assosiasjon mellom form og innhold, altså de er begge usualiserte, men

sistnevnte har åpenbart en langt høyere diffusjon på tvers av det norske språksamfunnet. Totalt sett har derfor *ord* en høyere konvensjonaliseringsgrad enn *leksem*.

Usualiseringa av sambandet mellom et uttrykk og et innhold kan i mange tilfeller knyttes til et bestemt domene. For eksempel er termen *leksem* først og fremst usualisert innenfor det språkvitenskapelige fagfeltet. Hvor allment tilgjengelig det aktuelle usualiseringsdomenet er, kan ha mye å si for hvor høy diffusjon uttrykket får. Om et ord usualiseres innenfor et domene som diskuteres mye i nasjonale medier, kan det føre til at det raskt får en høy diffusjon på tvers av språksamfunnet.

2.2.2 Innprenting

Fagtermen *innprenting* (entrenchment) lanseres gjennom Langackers (1987, 59) velkjente sitat: «every use of a structure has a positive impact on its degree of entrenchment (...)». Kort sagt betyr dette at hver gang en språkbruker produserer eller prosesserer et ord, en lyd, eller en større lingvistisk konstruksjon, så blir den lingvistiske enheten dypere innprenta i hen, hvilket seinere letter eller automatiserer prosesseringa av den aktuelle enheten. Innprenting er såleis en prosess som pågår på individnivå, og som påvirker den enkeltes begrepsapparat. Det er videre rimelig å anta at *n-gram* (altså ord eller ordsamband) som forekommer ofte i skrift og tale, har en høy sjanse for å bli innprenta hos de fleste av språksamfunnets medlemmer.

Zenner, Speelman og Geeraerts (2014) påpeker et viktig skille mellom kommunikativ og erfaringsbasert innprenting. Et begrep kan aktiveres mentalt gjennom kommunikative hendelser, altså at noen framstiller det språklig, men også gjennom rent fysiske situasjoner der man kommer i sansbar berøring med et fenomen som aktiverer begrepet, altså helt uten språklige stimuli. For eksempel kan vi komme til å få et mentalt bilde av kameler gjennom både å lese bokstavforbindelsen *kamel* og å ri på en faktisk kamel (mitt eksempel). Formodentlig er det bare den kommunikative innprentinga som bidrar til å styrke sambandet mellom en språklig form og begrepet den aktiverer.

Schmid (2020, 227) hevder at innprenting har to beslekta effekter: *rutinisering* (routinization) og *skjematisering* (schematization). Førstnevnte sikter til at repetisjon i produksjon og eksponering leder til at det gradvis blir en rein rutine for språkbrukeren å framkalle nødvendige assosiasjonsmønstre for prosessering av en gitt språklig enhet. Ved gjentatt framkalling av assosiasjonsmønstre knytta til et uttrykk som *bihulebetennelse* blir prosesseringa mer og mer en rein rutine, hvilket innebærer at det koster språkbrukeren mindre og mindre tid og krefter å framkalle de rette assosiasjonsmønstrene.

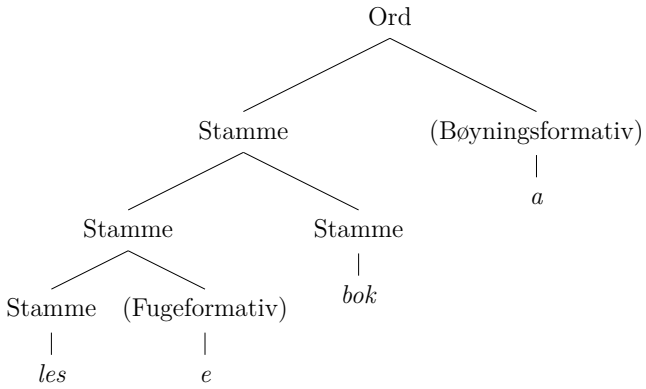
Skjematisering inngår som en integrert og uunngåelig effekt av rutiniseringsprosessen, og dreier seg om de mer subtile og overførbare generaliseringene som følger av rutinise-

ring (Schmid, 2020, 227–228). Samtidig som gjentakende eksponering og produksjon av sammensetninga *bihulebetennelse* (mitt eksempel) konsoliderer de helt spesifikke assosiasjonsmønstrene knytta til dette spesifikke uttrykket, konsolideres samtidig mer generelle, skjematiske egenskaper som *bihulebetennelse* utviser. På overordna nivå konsolideres regelen om at det er etterleddet som utgjør den semantiske og morfologiske kjernen i norske sammensetninger, mens på et lavere nivå konsolideres det produktive skjemaet der kroppsdel+betennelse betegner betennelser i ulike deler av kroppen. Skjematisering omfatter med andre ord den kognitive merverdien ved rutinisering av en konkret språklig konstruksjon.

I dette delkapittelet har jeg gjort greie for hvordan språkbruk er informasjonskilde til både lingvistisk og leksikografisk kunnskap. I det følgende vil jeg dreie fokuset mot sammensetninger mer spesifikt, og diskutere hva sammensetninger er, og hvilke variabler som påvirker deres status i objektet allmennleksikografien har som målsetning å beskrive.

2.3 Hva er en sammensetning?

Sammensetting er en produktiv orddanningsprosess som forekommer i omtrent samtlige av verdens språk (Pepper, 2020; Štekauer, Valera & Körtvélyessy, 2012). Det er en vanlig antakelse at sammensetninger består av to selvstendige ord (se blant annet Faarlund, Lie og Vannebo (1997) og Theil (2016)). Johannessen (2001) argumenterer likevel overbevisende for at sammensetninger mer presist består av to selvstendige *ordstammer*, det vil si den delen av et ord som holdes stabil på tvers av et bøyningsparadigme.[1] Analysen blir da at et sammensatt ord har følgende form (Johannessen, 2001, 73):



Til grunn for denne analysen ligger følgende premisser (hos Johannessen (2001), men liknende poeng fins hos Kinn og Kulbrandstad (2016, kap 7.1)):

[1] Selv om det fins godt om eksempler på stammer som endrer seg, for eksempel har ofte sterke verb vokalskifte i preteritum.

1. Forleddet i sammensetninger er ubøyd.
2. Eventuell sammensetningsfuge er et suffiks på forleddet.
3. Mange sammensetninger består av ledd som verken er selvstendige ord eller avledningsmorfer.
4. Bøyningsmorfer festes til den sammensatte stammen, ikke til etterleddet, selv om sammensetningas bøyningsparadigme som en hovedregel sammenfaller med etterleddets bøyningsparadigme.

Det finnes ingen overgripende konsensus om premissa ovenfor, og den norske sammensetningsfloraen er full av eksempler som utfordrer denne analysen. Dette har delvis å gjøre med at mange sammensetninger er levninger fra tidligere språksteg, og at mange er usystematisk importert fra andre språk. Dessuten behøver ikke en sammensetning være spesielt etablert før den er gjenstand for videre orddanning ved å inngå i nye sammensetninger, avledninger eller tilbakedanninger. Summen av dette blir en mangslungen sammensetningsflora som blant annet inneholder tilsynelatende bøyde forledd, som i *fedreland* og *tungtvann*, og ledd som ikke fungerer aleine, som etterledda i *rødøyd* og *venstrehendt*. Dessuten fins det mange grensetilfeller, spesielt i skillet mellom avledning og sammensetning. Noen illustrerende eksempler på dette trekkes fram i det følgende.

Om man setter to avledninger sammen, får man en sammensetning som i *ungdomskjærlighet* eller *nasjonalforsamling*. En forutsetning for dette er at de to ledda utgjør selvstendige ordstammer. På overflaten kan en avledning se helt lik ut som en sammensetning der etterleddet er en avledning. Avledningene *likegyldighet* og *tilbakeholdenhet* er til forveksling like sammensetningene *selvmedlidenhet* og *rettssikkerhet*. I mange tilfeller er det betydningsnivået som hjelper en å skille mellom de ulike konstruksjonene. For eksempel er *selvmedlidenhet* 'medlidenhet med seg selv', mens *likegyldighet* ikke er 'lik gyldighet', men 'det å være likegyldig'. Betydningsnivået kan såleis gi informasjon om hvor konstruksjonen lettest «knekker i to», altså mellom hvilke morfem eller morfer det gir mest mening å spalte den komplekse ordformen. Men i visse tilfeller kan det gi like god mening enten man spalter ordet på det ene eller andre stedet: *Arbeidsledighet* kan like gjerne parafraseres med 'ledighet fra arbeid' som 'det å være arbeidsledig'. Med andre ord kan konstruksjonen analyseres som både en sammensetning og en avledning.

Importord volder også problemer for grenseoppdraginga mellom avledning og sammensetning. Faarlund et al. (1997, 59) nevner at ord som *biologi* og *biografi* er sammensetninger i gresk. Selv om disse ledda ikke utgjør stammer for selvstendige ord i norsk, har også norske språkbrukere bevissthet om at slike ledd forekommer i mange ulike konstruksjoner i norsk (Kulbrandstad & Kinn, 2016, 7.1), og de kan dessuten knytte seg til ledd av ikke-gresk herkomst, f.eks. *biomangfold*, *biovitenskap*, *sexologi*. Jarvad (1995) betegner konstruksjoner med slike produktive, fremmede og ordliknende morfemer som *kryptosammensetninger*.

Johannessen (2001, 2017) og til dels Faarlund et al. (1997, 60) omtaler en annen konstruksjon som likner på, men ikke kan analyseres helt likt som sammensetninger og avledninger, nemlig *samdanninger*. Der referansegrammatikken slår seg til ro med at samdanninger i bunn og grunn er avleda sammensetninger, hevder Johannessen at samdanninger er en spesiell type sammensetninger der man har tre uunnværlige ledd. Om man fjerner et ledd fra samdanninga *rødøyd*, får man en av de ikke-etablerte ordformene **rødøye* eller **øyd*. Videre er samdanninger produktive sett bort fra begrensninga at det midtre leddet (f.eks. *øye* i *rødøyd*) må være et uavhendelig substantiv, typisk en kroppsdel. Alternativt kan man analysere *-øyd* som en bunden avleda stamme på lik linje med for eksempel *-holdig* i *blyholdig*.

Grensetilfella som trekkes fram i de tre seineste avsnitta, utgjør ikke en uttømmende liste over ordkonstruksjoner som utfordrer ordkategoriene simpleks, avledning og sammensetning.[2] Helst må nok disse kategoriene ansees å ha ulne grenser. Klassifiseringa av hver enkelt ordform informeres derfor snarere av forskningsspørsmålet enn av en absolutt sannhet om hvilken kategori hver ordform dypest sett tilhører. I leksikografisk sammenheng vil det være hensiktsmessig å ha i mente hva en typisk språkbruker er i stand til å identifisere. Uten å undervurdere den lingvistiske kompetansen i befolkninga er det ikke helt usannsynlig at en typisk språkbruker først og fremst vil bite seg merke i hva som fins på overflaten av et ord, og hva ordet betyr. Om formen til et ord inneholder bokstav- eller lydkombinasjonene som skal til for å tilstrekkelig minne om minst to mindre ord, vil nok de fleste oppfatte ordet som sammensatt. I visse tilfeller holder det faktisk at de kun gjenkjenner ett ledd. For eksempel behøver man ikke nødvendigvis å vite hva *bråte* eller *bring(e)* er for å anta at *bråtebrann* og *bringebær* er sammensetninger. Det holder altså med ett distinkt ledd, gitt selvfølgelig at man har leddbetydninger som passer noenlunde overens med den sammensatte betydninga. Om man kjenner betydninga til ordet *diskos*, vil man trolig ikke analysere konstruksjonen som en sammensetning av substantiva *disk* og *os*.

2.3.1 Definisjon av sammensatte ord

I denne avhandlinga anvender jeg sammensetningsdefinisjonen til Johannessen (2001) som sier at norske sammensetninger utgjøres av to ordstammer. For å avgjøre om noe er en ordstamme, er det nødvendig å undersøke om det som utgjør stammen, forekommer som selvstendig ordform enten i ordbøker eller språkbruk, eventuelt i kombinasjon med bøyningsmorfem. En forutsetning for at det er det samme morfemet som opptrer i de

[2] Blant annet kunne også *tilbakedanninger* av typen *sesongåpne* og *skreddersy* vært nevnt. Etymologisk er disse diskutable siden de er dannet av sammensetningene *sesongåpning* og *skreddersydd*, men om man ser bort fra etymologien og konsentrerer seg om de synkrone formene, kan slike dannelser som oftest segmenteres i to åpenbare og distinkte ordstammer.

ulike ordformene, at f.eks. etterleddet i *sjøsjuk* er simplekset *sjuk*, er at det er en tilstrekkelig betydningslikhet eller nær nok etymologisk forbindelse mellom de to morfema. I eksempelet *sjøsjuk* kan vi stadfeste betydningslikhet ved å konstatere at *sjøsjuk* er et hyponym til *sjuk*. I tilfellet *løvetann* er det ingen betydningslikhet mellom sammensetningas ledd og simpleksene *løve* og *tann*, men her fins det ifølge Det Norske Akademis Ordbok (heretter NAOB) (NAOB, 2023) en etymologisk kopling mellom simpleksenes betydning innafor og utafor sammensetninga.[3] En av grunnene til at enten betydningslikhet eller etymologisk kopling må være til stede, er at avledningsmorfem som simpelthen er homonyme med rotmorfem ikke skal forveksles, for eksempel er ikke suffikset i *sunnhet* relatert til adjektivet *het* 'varm'. Siden de leksikografiske kriteria for utvalg av sammensetninger i stor grad også vil kunne brukes for avledninger, innvilger jeg sammensetningsstatus i tvilstilfeller i denne studien.

Et typisk tvilstilfelle for definisjonen ovenfor skapes av morfem som villig inngår i komplekse ordformer, og som enten er homonyme med eller en forekomst av et simpleks eller en mindre kompleks ordform, der betydninga til morfemet i disse kontekstene er noe ulik, men likevel beslekta med betydninga til morfemet som ledd i komplekse ordformer. For eksempel er morfemet *-vis* oppført som både suffiks og simpleks i BOB. Dette taler for at komplekse ordformer med *-vis* er avledninger, som *nødvendigvis* og *naturligvis*. I disse eksempla er det åpenbart en viss avstand mellom simplekset *vis* 'måte' og suffikset *-vis* siden de komplekse formene tilhører andre ordklasser enn simplekset. Likevel vil en gjennomsnittlig språkbruker være i stand til å utskille tegn- eller lydsekvensen *vis* og gjenkjenne at denne korresponderer med simplekset *vis*.[4][5]

I de seineste avsnitta har jeg forsøkt å avgrense hva en sammensetning er, og hvordan sammensetninger kan defineres. I det følgende presenterer jeg en kort oversikt over sammensetninger som forskningsobjekt.

2.3.2 Kort forskningsoversikt

Mye forskning er gjort på sammensetninger, spesielt de siste to tiårene. Internasjonalt har det utkommet i alle fall fire antologier i denne perioden. Antologien til Lieber og Štekauer (2009) samler semantiske og typologiske tilnærminger til å beskrive sammensatte ord, mens antologien til Štekauer et al. (2012) samler typologiske tilnærminger til ulike orddanningsprosesser, deriblant sammensetting, mens en annen antologi, Scalise og Vogel (2010), har en bredere tverrfaglig innfallsvinkel til beskrivelse av sammensetninger. I

[3] Navnet *løvetann* er et oversettingslån fra gresk *leontodon* som angivelig er motivert av 'bladflikenes likhet med rovdyrtenner'

[4] NAOB oppgir dessuten at avledningsmorfemet *-vis* er «samme ord» som substantivet *vis*.

[5] Derfor innvilges for eksempel *eksempelvis* sammensetningsstatus i artikkel 3, uten at dette er ment å forfekte noen allmenngyldig sammensetningsdefinisjon.

denne antologien diskuteres blant annet syntaktiske, morfologiske, fonologiske, semantiske, typologiske, kvantitative, psykolingvistiske og tegnspråklige problemstillinger knytta til sammensetting og sammensetninger. I tillegg fins antologien til ten Hacken (2016), som samler ulike innfallsvinkler til å beskrive semantikken til ulike sammensetningstyper i ulike språk. I skandinavisk sammenheng er det på sin plass å nevne følgende verk (i kronologisk rekkefølge) fra 1990 og framover:

- Leira (1992) gir en grundig og rikholdig oversikt over ordlagingsmekanismene sammensetting og avleding i bokmål og nynorsk, og forklarer dessuten sentrale mønstre ved ulike sammensetningstyper. I tillegg bruker han mange eksempler i sin redegjørelse av fenomen som innkorting av forledd, sammensetningsfuger og tilpassing av lånord. Kort sagt kan man si at Leiras verk er en bortimot uttømmende typologi over norske sammensetninger, riktignok med et nesten ensidig fokus på morfologi.

- *Norsk referansegrammatikk* (Faarlund et al., 1997) har også en utfyllende sammensetningstypologi for norsk, mye i samme stil som Leira (1992). De to oversiktene har mye til felles, blant annet at de gir detaljerte beskrivelser av ulike ordklassekomposisjoner i sammensatte ord, og at de anerkjenner den variable semantikken i sammensetninger. En forskjell er likevel at referansegrammatikken også gjør en semantisk inndeling av sammensatte ord.

- Mellenius (1997) bruker i sin avhandling en mangefasettert innfallsvinkel til å studere hvordan svenske barn tilegner seg evnen til å produsere nylagde sammensetninger (novel compounds). Studien er dels basert på ulike eksperimenter og dels på forfatterens observasjoner av egne barn.

- Bakken (1998) tar i sin avhandling utgangspunkt i et gammelnorsk diplommateriale for å skildre leksikaliseringsprosessen som blant annet norske sammensetninger gjennomgår (se 2.5 for mer detaljert omtale av Bakkens avhandling).

- Sakshaug (1999) setter i sin avhandling særskilt søkelys på sammensetninger med verbalsubstantiver (deverbal nouns), av typen *brøkrekning*, *skihopp* eller *kyllingklekkeri*, og foreslår en modulbasert analyse av disse der «the analyses of morphology, syntax, and semantics are given a chance on their own» (212).

- Johannessen (2001) gir en drøfting av hvorvidt sammensetninger er en komposisjon av ord eller stammer. Hun imøtegår ordleddsanalysen som fins blant annet hos Faarlund et al. (1997), og argumenterer for at en stammeleddsanalyse gir langt mer mening i lys av språklig empiri.

- Svanlund (2002) betrakter sammensetningers semantiske gjennomsiktighet fra et kognitivlingvistisk perspektiv, og kommer til dels med kritikk av leksikaliseringsbegrepet som fins hos blant andre Bakken (1998) (se 2.5.4 for nærmere omtale

av denne kritikken til Svanlund (2002)). Videre foretar Svanlund (2009) også en grundig korpusundersøkelse av den «leksikalske etableringa» til et knippe relativt nylagde svenske sammensetninger, dvs. hvordan disse sammensetningenes form–betydning-par har oppstått og blitt konvensjonelle.

- Bakken og Vikør (2011) vender oppmerksomheten mot en lite beakta gruppe sammensetninger, nemlig sammensatte preposisjoner, og gjør en teoretisk fortolkning av prosessen slike preposisjoner gjennomgår i talemålet fra usammensatt form til sammensatt og fusjonert form (for eksempel at den norrøne preposisjonen *á* har utvikla seg via *uppá* og *uppå* til den synkrone *på*, som igjen inngår i nye preposisjonssamband).

- Nesset (2011; 2017; 2018; 2019) er sannsynligvis den som har skrevet mest om norske sammensetninger fra et kognitivlingvistisk perspektiv. I fire artikler analyserer han ulike norske sammensetningstyper, blant annet personkarakteriserende sammensetninger, som *skravlebøtte* og *treneve*, og kulturelt betinga sammensetninger, som *makrellfotball* og *lofotfiske*. Sentralt i analysene står begrepa metaforikk og metonymi (etter Lakoff og Johnson (1984)) og konseptuell integrasjon (etter Fauconnier og Turner (2008)).

- Eiesland (2015) foretar i sin avhandling en kategorisering av rundt 2000 norske substantiv-substantiv-sammensetninger basert på den semantiske relasjonen mellom ledda. For eksempel klassifiseres relasjonen i sammensetningene *diamantring* og *eplekake* som karakteristisk del (characteristic part).

- Theil (2016) formulerer en rekke fonologiske, morfologiske og semantiske kriterier han hevder gjelder for prototypiske norske sammensetninger.

- Kjelsvik (2017) gjør blant annet rede for hvordan både prototypiske sammensetninger og grensetilfeller av sammensetninger har blitt håndtert i standardordbøkene.

- Eik (2019) formulerer i sin avhandling en fyldig morfosyntaktisk analyse av norske sammensetninger. Analysen er fundert på et omfattende teorigrunnlag og munner ut i et forslag til en overordna morfosyntaktisk trestruktur over sammensetningers universelle komponenter.

- Loenheim (2019) undersøker i sin avhandling fortolkninger av en variert gruppe sammensatte ord i svensk. Tolkningene er gjort av elever i videregående skole hvorav omkring halvparten har svensk som førstespråk, mens resten har svensk som andrespråk.

En kjensgjerning en kan utlede fra forskninga som er nevnt ovenfor, i tillegg til forskning jeg vil presentere i det følgende, er at sammensetninger har ulik status i henhold til en

rekke variabler. I det følgende vil jeg redegjøre for og diskutere de tre variablene som spiller størst rolle i denne avhandlinga, nemlig etablering, anomalisering og utbredelse.

2.4 Etablering

I dette delkapittelet forklares det hvorfor og på hvilken måte sammensetninger har ulik status med hensyn til *etableringsgrad*. Sammensetningsforskninga er prega av forskjelligarta terminologi for til dels overlappende og ekvivalente fenomen. I denne avhandlinga vil jeg følge samme terminologiske konvensjoner som Svanlund (2009) med hensyn til å benevne sammensetningers kognitive og sosiale status i språket. *Etablering* brukes her for å betegne hvor innarbeida sammensetninger er i språket generelt sett og på tvers av alle dets brukere. Etablering er såleis summen av konvensjonalisering og innprenting (se underkapitla 2.2.1 og 2.2.2). Når en sammensetning er veletablert, betyr dette at den både er innprenta i språkbrukernes begrepsapparat og akseptert i språksamfunnet som en konvensjonell betegnelse. Her må det understrekes at etableringa knytter seg til sammenhengen mellom en lydlig eller ortografisk form, et eller flere korresponderende mentale begrep, og et eller flere korresponderende fenomen.

Svanlund (2002) påpeker at etableringa av sammenhengen mellom en form, et begrep og et fenomen kan ses fra både en semasiologisk og en onomasiologisk synsvinkel. Førstnevnte dreier seg om hvor etablert betydning et uttrykk har. Sistnevnte dreier seg om hvor etablert uttrykk et begrep har. I det følgende vil jeg kaste lys over hvordan sammensetninger har ulik etableringsgrad sett fra disse ulike, men overlappende synsvinklene.

Svanlund (2009, 20) gir med utgangspunkt i sammensetningslitteraturen en sammenfattende liste over hva sammensetninger typisk brukes til:

1. navngi
2. betegne underkategorier til mer generelle kategorier
3. fylle ut mangler i det allmenne leksikonet (språksamfunnet savner betegnelse for begrepet)
4. kompensere for mangler i det private leksikonet (taleren kan ikke det konvensjonelle ordet eller kommer ikke på det)
5. markere kontraster
6. muliggjøre mer presis kommunikasjon
7. gi en konsentrert framstilling
8. hentyde anaforisk til noe som har blitt mentalt aktivert tidligere i konteksten

Mens punkt 4–8 viser til anvendelsen av sammensetninger innenfor reint kontekstuelle formål, viser punkt 1–3 til mer varige formål. Likevel er det fullt mulig at sammensetninger som initielt benyttes til alle de ovenstående formålene, kan etablere seg som

den primære betegnelsen på et fenomen. Denne mekanismen har fått relativt sparsomt med oppmerksomhet, men blant annet Mellenius (1997) kommenterer kort at en sammensetning som blir etablert, mottar egenskaper og betydningsnyanser fra fenomenet den denoterer. På samme måte som et person- eller stedsnavn er uløselig knytta til henholdsvis en person eller et sted, så blir en sammensetning uløselig knytta til sitt korresponderende fenomen. Eventuelle fysiske, utviklingsmessige eller assosiative endringer ved fenomenet vil da også påvirke de mentale forestillingene sammensetninga vekker. Et godt eksempel på at et navns valør og fargenyanser kan omskapes over tid er personnavnet *Jeanne d'Arc*. At personen dette navnet refererer til, i sin samtid ble brent på bålet for hekseri, for så omtrent 500 år seinere å bli kanonisert av paven (Wikipedia 2023), sier noe om at det konnotative innholdet i navnet må ha endra seg gjennom tidens løp, selv om referansen holder seg stabil. Nesten tilsvarende, men med en viktig forskjell, har det konnotative innholdet i *datamaskin* endra seg over tid. Forskjellen her er at også denotasjonen til *datamaskin* har endra seg. I takt med teknologisk utvikling har fenomenet som *datamaskin* korresponderer med, blitt mer mangfoldig og avansert. Selv om en datamaskin som bygges i 2023, har vesentlig andre egenskaper enn de tidligste datamaskinene, så tjener like fullt *datamaskin* som betegnelse på dagens fenomen.

Når sammensetninga *datamaskin* har en navneliknende funksjon, dreier dette seg om en viss etableringsgrad på både semasiologisk og onomasiologisk nivå – uttrykket viser relativt konsekvent til samme fenomen og fenomenet blir relativt konsekvent betegna med uttrykket *datamaskin*. Dette står ikke i veien for at uttrykket også kan benyttes om andre fenomen, enten via metafor og metonymi, eller ved at en potensiell, ikke-konvensjonalisert betydning blir aktivert. Nesset (2017, 92) påpeker for eksempel at den konvensjonaliserte sammensetninga *lofotfiske*, som vanligvis betegner 'vinterfiske etter skrei i Lofoten', også kan anvendes med den potensielle betydninga 'fiske i Lofoten'. Dette henger sammen med noe Loenheim (2019, 50) og Svanlund (2002) understreker, nemlig at sammensetninger er grunnleggende polysemiske størrelser, og at de dermed har et stort semasiologisk betydningspotensial. Konvensjonaliseringsprosesser bidrar imidlertid til at visse betydninger befester seg hos språkbrukerne slik at andre like plausible betydninger blir mindre aktuelle. Ikke-konvensjonelle sammensetninger tolkes på sin side med utgangspunkt i språklige skjema på ulike abstraksjonsnivå, for eksempel med utgangspunkt i analoge etablerte sammensetninger eller mer abstrakte kognitive skjema. Dette viser at språkbrukere vil forsøke å tolke et hvilket som helst uttrykk med den kunnskapen de har tilgjengelig, og med utgangspunkt i hva som kan tenkes å være relevant innenfor den språklige konteksten. Om de er tidligere eksponert for uttrykket, med andre ord at uttrykket har en viss semasiologisk innprenting, kan de dra nytte av sitt språklige minne, men om de har lite eller ingen tidligere eksponering, vil de sannsynligvis finne andre utgangspunkt for tolkninga. Dette betyr ikke at det språklige minnet er helt uvirksomt i tolkninga av ukjente sammensetninger, snarere må minnet aktiveres for

å tolke sammensetningsledda, samtidig som det anvendes for å finne analoge etablerte sammensetninger.

Med en onomasiologisk tilnærming tar man altså, til forskjell fra en semasiologisk tilnærming, utgangspunkt i det betegna, det vil si begrep, og undersøker kort sagt hvordan dette blir aktivert av ulike uttrykk (Geeraerts, 2006). Den onomasiologiske etableringsgraden til en sammensetning blir i dette bildet med hvilken relative hyppighet, gitt alle ganger et begrep aktiveres (Mehl, 2016, 48), den aktuelle sammensetninga blir benytta. For eksempel kan man ta utgangspunkt i begrepet BUKSE og dets mulige betegnelser, f.eks. *bukse*, *benklær*, *brok*. Dersom et uttrykk er klart foretrukket foran alternativa, kan man slå fast at dette uttrykket har sterkest onomasiologisk etableringsgrad for det aktuelle begrepet, eller høyest *onomasiologisk saliens* i termene til Geeraerts (2000). Visse uttrykk kan være så godt som enerådende betegnelser for sine fenomen. For eksempel er *alkoholbriller* i prinsippet dekkende for innholdet som betegnes av *ølbriller*, altså 'sviktende vurderingsevne grunnet alkoholinntak' (NAOB). Likevel har førstnevnte ingen treff i Nasjonalbibliotekets n-gramsøk. Her blir muligens ordformen *alkoholbriller* blokkert av at *ølbriller* tjener som en navneliknende betegnelse på fenomenet.

Mindre etablert variant	Mer etablert(e) variant(er)
melkenaut	*melkeku*
fåreull	*saueull*
grisekotelett	*svinekotelett*
gampehale	*hestehale*
badetruse	*badebukse*, *bikini(underdel)*, *speedo*
lærvest	*skinnvest*
knallvakker	*smellvakker*, *bråvakker*
snikrøyke	*smugrøyke*

Tabell 2.1: Onomasiologiske varianter med ulik etableringsgrad

Felles for sammensetningene i venstre kolonne i tabell 2.1 er at de har en relativt åpenbar referanse, hvilket vil si at de har en tilstrekkelig semasiologisk forutsigbarhet til å være gangbare uttrykk med en uproblematisk definisjon. Onomasiologisk er de likevel ikke tilstrekkelig foretrukne betegnelser for sine referenter til å kunne sies å være etablerte uttrykk. De er med andre ord et godt stykke unna å tjene som betegnelse på lik linje med de tilsvarende etablerte sammensetningene.

I leksikografisk sammenheng er både semasiologisk og onomasiologisk etableringsgrad relevant. På den ene siden må eller bør oppslagsord ha minst én etablert betydning, hvis ikke hadde det vært fånyttes å gi dem en tilfredsstillende definisjon. På den annen side bør oppslagsorda være etablerte uttrykk for fenomenene de betegner. Som vi vil se i neste delkapittel, har nylagde, uetablerte og gjennomsiktige sammensetninger ofte via sine ledd og komposisjonelle mønster en tilstrekkelig semasiologisk forutsigbarhet til

å være fullt gangbare uttrykk innenfor mange kontekster, uten at de har tilstrekkelig onomasiologisk etableringsgrad til å være gunstige ordbokkandidater.

Downings distinksjon mellom «name-worthy categories» og «name-worthy entities» er dessuten aktuell i leksikografisk sammenheng (1977, 823). Enkelte sammensetninger, for eksempel *apple-juice seat*, fungerer bare innenfor en smal og spesifikk kontekst, og brukes derfor til rent deiktiske, dvs. påpekende formål (se dessuten punkt 4-8 i innledninga av dette delkapittelet). Om man har en skuff full av votter og en annen full av luer, vil det gi god mening å benevne disse *votteskuffen* og *lueskuffen*, men man kan ikke godt dra til nærmeste møbelforretning og be om å få se assortimentet deres av votte- og lueskuffer. Sammensetningenes påpekende funksjon står seg bare innenfor et veldig begrensa sett av kontekster; de denoterer altså ikke en klasse av objekter, men viser helt enkelt til hvert sitt konkrete objekt som assosieres med det innholdet de innenfor et avgrensa tidsrom rommer.[6] Downing spesifiserer at deiktiske sammensetninger som *apple-juice seat* og *votteskuff* vitner om en distinksjon mellom «name-worthy categories» og «name-worthy entities», hvorav de deiktiske sammensetningene er eksempler på det siste. Vi bruker med andre ord sammensetninger til å betegne både umiddelbare, men forgjengelige objekter og fenomener, og varige kategorier.

Selv om de fleste ordbøker har ordtilfanget som utgangspunkt, må leksikografer også til en viss grad vurdere hva slags fenomentilfang ordboka skal fange. En viktig skillelinje her er hvorvidt en sammensetning betegner en kategori eller en (midlertidig) entitet. For selv om høy grad av navnelikhet gjennom etablering i prinsippet øker ordbokaktualiteten til en sammensetning, må ikke navnelikheten overskride skjæringspunktet mellom proprium og appellativ – et skille som nærmest kan betegnes som forskjellen på entall og flertall. Der *Frihetsgudinnen* blir et proprium fordi det betegner én entitet,[7], er *amorin* et appellativ fordi det betegner en klasse av entiteter (et flertall statuer og bilder) og derved en kategori. Som vi har sett, betegner også en del svært kontekstbundne sammensetninger som *votteskuff* også bare én referent innenfor den aktuelle konteksten.

I dette delkapittelet har jeg redegjort for hvordan sammensetninger i ulik grad er semasiologisk og onomasiologisk etablerte. Som et resultat av konvensjonalisering og innprenting kan det dannes uløselige assosiative tilknytninger mellom uttrykk og innhold. Men siden både konvensjonalisering og innprenting er kontinuerlige prosesser, vil naturligvis visse assosiative koplinger være løsere enn andre, hvilket gjør noen sammensetninger til bedre ordbokkandidater enn andre.

[6] Det er imidlertid ingenting i veien for at noen kan spesialdesigne en votteskuff og slik skape en mer stabil denotasjon for dette uttrykket. Det fins dessuten en ikke ubetydelig mengde belegg på sammensetninga *sokkeskuff* i Nasjonalbibliotekets n-gramssøk, noe som kan indikere at KLESPLAGG + *skuff*-skjemaet kan anvendes til noe mer enn bare rent deiktiske formål.

[7] Selv om det fins flere like statuer (gjerne omtalt som kopier) blant annet i Paris og på Karmøy.

2.5 Anomalisering

Et hovedspor innenfor sammensetningslitteraturen kan kalles *leksikaliseringssporet*. På dette sporet postuleres det at ord gjennomgår en prosess der de over tid utvikler det som i synkron forstand kan forstås som semantiske, grammatiske eller til og med fonologiske uregelmessigheter (se for eksempel Bakken (1998, 2006); Bauer (1983); Downing (1977); Eik (2019); Libben (1998); Libben, Gibson, Yoon og Sandra (2003). Innenfor dette rammeverket tjener utviklinga av de uregelmessige trekka som en indikasjon på økt etablering. Et sentralt verk innenfor leksikaliseringssporet i norsk sammenheng er Bakken (1998) sin studie av den diakrone prosessen som (ofte) blir sammensatte ord til del. Med hennes termer er nylagde sammensetninger ansett å være *motiverte* og *komposisjonelle*, mens sammensetninger som fullt ut har gjennomgått en lang leksikaliseringsprosess, og som dermed befinner seg i den mest ugjennomsiktige enden av leksikaliseringsskalaen, benevnes som *demotiverte* og *ikke-komposisjonelle*. Leksikaliseringsprosessen innebærer ifølge Bakken en konvensjonalisering av sambandet mellom et spesifikt uttrykk og dets innhold til den grad at sambandet etter hvert framstår som arbitrært på lik linje med sambandet mellom simplekser og deres innhold: «Sammensetningen får dermed mer og mer karakter av å være et minimalt språklig tegn» (Bakken, 1998, 61). Sammensetninger som gjennomgår denne prosessen, mister ifølge Bakken sin interne morfologiske struktur.

I denne avhandlinga oversetter jeg leksikaliseringstermen til henholdsvis *konvensjonalisering*, *innprenting*, *etablering* eller *anomalisering*, alt etter om det refererte verket betegner en sosial, kognitiv, eller overgripende prosess, eller utviklinga av uregelmessige trekk. Tidvis bruker for eksempel Bakken *leksikalisering* om en prosess som leder til økt konvensjonalitet, og såleis kan den erstattes av *konvensjonalisering*. Et argument for å fase ut leksikaliseringstermen er at den viderefører ideen om at språkbrukeres ordkunnskap er delt mellom to avgrensa moduler, et mentalt leksikon og en grammatikk. I dette bildet forutsettes det at *komposisjonelle* sammensetninger (se 2.5.2) tilhører grammatikken fordi språkbrukere kan gjøre spontane, regelbundne tolkninger av disse sammensetningenes betydning, mens sammensetninger med uregelbunden betydning må lagres enkeltvis i språkbrukerens minne, kalt leksikon. En slik begrepsliggjøring av ordkunnskap bygger ifølge Svanlund (2002) på et forenkla syn på ordsemantikk, og overser hvordan assosiasjonsmønstre mellom språklige former og innhold kan endres over tid i takt med ekstralingvistiske forandringer i verden, eller gjennom suksessiv språkbruk. Kort sagt gjennom innprenting eller konvensjonalisering.

Downing (1977, 820) poengterer at det ikke først og fremst er etablerte ordformer som byr på utfordringer for synkrone analyser av sammensetninger. Snarere er det deres forbindelse til «real world change». En sammensetning kan være fullt ut motivert og gjennomsiktig ved sin tilblivelse, men så snart den har blitt innlemma av språksamfunnet

som en konvensjonalisert betegnelse, kan formen framstå like arbitrær som et hvilket som helst monomorfemisk ord, jf. følgende sitat av Gardiner (1954, 20):

> It may be said, indeed, that we must have had some reason for giving them those names rather than any others; and this is true; but the name, once given, is independent of the reason ... a town may have been named Dartmouth, because it is situated at the mouth of the Dart. But ... if sand should choke up the mouth of the river, or an earthquake change its course, and remove it to a distance from the town, the name of the town would not necessarily be changed.

Med andre ord behøver bare uttrykket å være gjennomsiktig i det øyeblikket det først ble ytra eller innlemma i språksamfunnet (se f.eks. Ryder (1989) og Zimmer (1971, 1972) for liknende poenger).

I det følgende vil jeg bruke Svanlunds term *anomalisering* for å betegne semantiske eller formelle uregelmessigheter knytta til sammensatte ord. Semantisk eller formell anomali kan, som Bakken (1998), Eik (2019) og Svanlund (2002) påpeker, være et symptom på høy etableringsgrad. Svanlund (2002) bruker termen *anomalisering* for å betegne en tilstand der flermorfemiske uttrykk har en annen betydning enn det som (på overflaten) framgår av de inngående morfema. Selv om termen *anomalisering* på samme måte som *leksikalisering* antyder en prosess fra ikke-anomal til anomal, påpeker Svanlund at selv helt nylagde sammensetninger fint kan være anomale. Spørsmålet om semantisk anomali åpner en diskusjon om hva det vil si for en sammensetning å være komposisjonell og motivert. Videre forutsetter termen *anomali* en klar oppfatning om anomaliens motstykke, nemlig det regelmessige. Selv om det ikke fins noen konsensus om hvordan man sporer anomali i sammensetninger, presenteres noen sentrale innfallsvinkler til å beskrive hva som er typisk og atypisk for norske sammensetninger i underkapitlene 2.5.1 nedenfor.

Anomaliseringsgrad er et mulig, men ikke nødvendig symptom på etableringsgrad, og kan knyttes til både formelle, herunder morfofonologiske, og semantiske egenskaper ved sammensetninger. Etablerte sammensetninger kan være uten anomali, samtidig som relativt uetablerte sammensetninger kan være anomale. Selv om anomali i siste instans ikke er et direkte symptom på etableringsgrad, har det en verdi i seg selv å inkludere anomale sammensetninger i ordbøker, nettopp fordi deres semantiske eller morfologiske egenskaper gjør dem iøynefallende eller uforutsigbare for den jevne språkbruker, hvilket presumptivt skaper et brukerbehov. At anomalisering ikke er en nødvendig følge av etableringsgrad, eksemplifiseres av Svanlund (2002) med sammensetninga *högintressant*. Selv om det verken fins noen semantisk eller formell anomali knytta til denne sammensetninga, har den en onomasiologisk etableringsgrad som gjør at den foretrekkes framfor

prinsipielt jevngode alternativer som *storintressant*. Dette viser at sammensetninger kan konvensjonaliseres og innprentes uten å utvikle avvikende trekk.

2.5.1 Hva er regelmessig?

For å kunne identifisere anomalier må man ha et begrep om hva som ikke er anomali, altså hva som er regelmessig. Som nevnt i delkapittel 2.3 og av Theil (2016) følger prototypiske sammensetninger samme bøyningsparadigme som etterleddet. Sammensetninger hvis bøyning skiller seg fra etterleddets bøyningsparadigme utenfor sammensetningskonteksten, må derfor anses for å være formelt avvikende. Semantisk spesifiserer Theil at mindre prototypiske sammensetninger kan avvike ved å inneha varierende mønster, manglende hode eller avvikende betydning (2016, 245). Etterleddet *mann* har for eksempel varierende mønster i de tre sammensetningene *bymann*, *yngstemann* og *sistemann*. Der *bymann* refererer til en voksen person av hankjønn, refererer*yngstemann* til en person (sannsynligvis hankjønn) som er yngst i en flokk uavhengig av egen alder, mens *sistemann* refererer til en person som kommer sist, uavhengig av kjønn og alder.

Med «avvikende betydning» sikter Theil til sammensetninger med sammensetningsledd som betegner fenomener som ikke kan sies å tilhøre den prototypiske kategorien TING. Selv om Theil (2016) helt klart har rett i at substantiv-substantiv-sammensetninger er den mest produktive sammensetningstypen i norsk (se for øvrig Paulsen (2023)), står man i fare for å vanne ut anomali-begrepet om man slår seg til ro med at alle andre sammensetningstyper er anomalier. Sammensettingsfunksjonen er tross alt produktiv på tvers av ordklasser i norsk, og sammensetninger som *ansvarliggjøre* og *fordomsfri* (mine eksempler) bør kunne sies å være temmelig ordinære, i hvert fall semantisk sett.

Theil (2016) nevner ikke komposisjonalitet eller semantisk gjennomsiktighet i sin beskrivelse av prototypiske sammensetninger, men dette tar ellers stor plass i sammensetningslitteraturen. Både Leira (1992, 21), Faarlund et al. (1997, kap. 2) og Fjeld og Vikør (2008, 46–49) nevner at forleddet i en sammensetning typisk innskrenker betydninga til etterleddet, men at det på ulike måter fins visse unntak fra denne regelen. Et eksempel på unntak er sammensetninger som over tid har utvikla avvikende betydninger, og som derfor ikke helt enkelt kan tolkes med utgangspunkt i ledda aleine. Sammensetningene *grusvei*, *kappspringe*, *purpurrød* (mine eksempler) har alle forledd som bidrar til å gjøre sammensetningas betydning til en innsnevra versjon av etterleddets selvstendige betydning. At fenomenet disse sammensetningene betegner, er en undertype av etterleddets korresponderende fenomen, kan blant annet understrekes ved at etterleddet kan brukes aleine for å referere til mer eller mindre samme fenomen eller entitet, gitt en etablert kontekst, se eksempel 1–3.

(1) De gikk langs en **grusvei**. **Veien** var svingete og smal.

(2) Barna **kappsprang** fram til lyktestolpen. Så gikk de tilbake til treet og **sprang** til lyktestolpen igjen.

(3) Konduktøren hadde en blå lue og et **purpurrødt** skjerf. Det **røde** utgjorde en flott kontrast til det blå.

Det samme lar seg ikke gjøre dersom den sammensatte betydninga ikke er en innsnevra versjon av etterleddets selvstendige betydning.

(4) *Smårollingen plukket en **løvetann**. **Tanna** hadde en lang stilk.

(5) *Jeg **overså** et problem i går. Og i dag **så** jeg nok et problem.

(6) *De var **likeglade** med resultatet. Akkurat som de var **glade** over innsatsen.

Siden *løvetann* ikke denoterer en type tann, *overse* ikke denoterer en måte å se på og *likeglad* ikke denoterer en undertype av det å være i godt humør, kan ikke etterledda brukes aleine for å henvise til samme referent eller fenomen som sammensetningene. Blant andre Bakken (1998, 62) og Eik (2019, 227) vil hevde at sammensetningene i den siste gruppa har simplekse betydninger og fungerer på samme måte som arbitrære simpleks. Det er umiddelbart fristende å hevde at sammensetningene i denne gruppa har en avvikende semantikk, men det ville vært forskningshistorieløst å ta endelig stilling til hva som er regelmessig og ikke for sammensetningers semantikk uten å komme inn på begrepa KOMPOSISJONALITET, GJENNOMSIKTIGHET og MOTIVASJON. Dette gjøres derfor i det følgende.

2.5.2 Komposisjonalitet

Semantiske anomalier kan som nevnt knyttes til sammensetningers gjennomsiktighet, komposisjonalitet og motivasjon. Langacker (1987, 448-480) redegjør for sammensetninger via en prinsipiell og overordna diskusjon om ulike egenskaper ved sammensetting som språklig fenomen. Denne diskusjonen begrenser seg ikke til sammensetting som orddanningsfunksjon, men omfatter også enhver grammatisk konstruksjon som kan sies å bestå av flere komponenter.[8] Han skiller mellom komplekse konstruksjoners *analyserbarhet* og *komposisjonalitet*. Førstnevnte dreier seg om hvorvidt det lar seg gjøre å gjenkjenne hva hver enkelt komponent i en kompleks konstruksjon bidrar med til helheten. Analyserbarhet er altså et spørsmål om hvorvidt for eksempel en helhetsbetydning av en sammensetning lar seg bryte opp i ulike deler som deretter kan tilskrives de ulike sammensetningsledda. Komposisjonalitet er derimot, ifølge Langacker (1987, 448), «the

[8]For eksempel består en setning av setningsledd, setningsledd består av ord, mens ord kan bestå av mindre ord.

degree to which the value of the whole is predictable from the value of its parts». Med andre ord er spørsmålet om komposisjonalitet direkte relatert til språkbrukernes evne til å produsere og fortolke uetablerte sammensetninger. Langacker hevder at komplekse språklige konstruksjoner som en hovedregel ikke kan tilskrives full komposisjonalitet. Dette av den enkle grunn at komplekse konstruksjoner ofte rommer spesifikasjoner som ikke kan spores til deres bestanddeler. Det er ingenting i sammensetninga *båthus* (mitt eksempel) som forteller oss om det skal forstås som 'hus på båt', 'hus til båt' eller 'hus som også er en båt'. Relasjonen mellom ledda er underspesifisert. I det hele tatt advarer Langacker mot å begrepsliggjøre lingvistiske komposisjoner via en *byggeklossmetafor* (building-block metaphor) der komponentene svarer til byggeklosser som må stables på en bestemt måte for at helhetsverdien (eller -betydninga) av komposisjonen skal åpenbare seg. I stedet argumenterer Langacker for å anskue språklige komponenter, som for eksempel sammensetningsledd, som tilganger til ulike kunnskapssystemer. I dette bildet er ikke sammensetningsledda stabile og fast avgrensa byggeklosser, men snarere en type nøkler som gir tilgang til ulike domener. Formodentlig ville Langacker hevda at *løvetann* (mitt eksempel) verken er analyserbart eller komposisjonelt. Nøklene *løve* og *tann* aktiverer mentale begrep som må sies å være temmelig irrelevante for å komme fram til den botaniske betydninga av *løvetann*.

Komposisjonelle sammensetninger er i henhold til Langacker (1987) forutsigbare fordi de aktiverer regulære, skjematiske strukturer hos språkbrukeren som forteller vedkommende hvordan sammensetningas betydningselementer, eller dens kombinasjon av kunnskapssystemer, skal fortolkes. Selv sammensetninger som komprimerer omfattende betydningsinnhold, og som dermed framstår som idiosynkratiske, som for eksempel *blackbird*[9] 'svarttrost' (1987, 450), bygger ifølge Langacker på en skjematisk struktur. For eksempel følger den hovedregelen om at etterleddet utgjør det semantiske hodet. Sånn sett er også *blackbird* til en viss grad komposisjonelt. Langacker hevder at det vil være en feiltakelse å slå seg til ro med at sammensetninger av denne typen simpelthen er ugjennomsiktige oppføringer i språkbrukernes leksikon som er ubundne av skjematiske grammatiske regler.

Bundgaard, Ostergaard og Stjernfelt (2006) utdyper Langackers forestillinger om komposisjonalitet via Fauconnier og Turners (2008) begrep om konseptuell integrasjon. Sammensetninger er komposisjonelle fordi de bygger på en konstruksjon XY som aktiverer et skjema der Y utgjør en ramme som X spesifiserer videre. Sammensetninger aktiverer ikke nødvendigvis det som i Fauconnier og Turners termer kalles *blandingsrom* (blends) der en «framvoksende struktur» (emergent structure) oppstår i integrasjonen mellom X og Y. Snarere hevder Bundgaard et al. (2006) at sammensetningskonstruksjonen er

[9]Langacker særskriver denne sammensetninga, mens blant annet Merriam-Webster oppgir at den skal samskrives.

asymmetrisk fordi ledda i ulik grad bidrar til å skape betydning. Y oppretter en skjematisk ramme, og X spesifiserer en av denne rammens soner eller felter (slots). Likevel er denne relasjonen høyst uforutsigbar og tidvis umulig å fastslå. Listen over mulige relasjoner mellom XY er, om ikke uendelig, så i alle fall veldig lang og fleksibel (jf. Eiesland (2015)). For eksempel er det uklart hvilket felt forleddet spesifiserer i en sammensetning som *rain forest* 'regnskog' siden 'skog der det regner mye' trolig ikke er en fullgod parafrasering. Det virker dessuten søkt at etterleddet *skog* aktiverer en ramme som har et spesifikt attributt VÆRTYPE. Det er også urimelig å hevde at tilstedeværelsen av regn er det som skiller regnskoger fra andre skoger, selv om førstnevnte helt klart har en løs assosiativ kopling til regnfall. Med dette ser vi at den semantiske informasjonen eller spesifiseringa som X gjør av Y kan være forholdsvis vag. Et annet eksempel er sammensetninga *sykkelpumpe* (mitt eksempel), hvor det fins minst to mulige analyser. Én analyse er at forleddet *sykkel* inngår i en HELHET FOR DEL-metonymi der den reelle referenten er sykkelslangen.[10] En annen analyse er at forleddet simpelthen angir et domene, eller en aktiv sone (se Langacker (1987, 273) som referenten til etterleddet hører inn under. Siden det vanligvis bare er én bestanddel som behøver pumping på en sykkel, er den uspesifikke hentydinga til domenet SYKKEL tilstrekkelig for at sammensetninga skal få en adekvat fortolkning.

Bundgaard et al. (2006) åpner likevel for at enkelte sammensetninger, som *mall rat* (kjøpesenterrotte), aktiverer blandingsrom. Nesset (2017) framhever det samme poenget via de norske sammensetningene *makrellfotball* (som betegner en viss type angrepsfotball fra fotballaget Start) og *traktoregg* (som betegner en rundballe). Typisk for slike sammensetninger er at det fins en begrepslig avstand mellom den konvensjonelle betydninga til ledda og sammensetninga som helhet, for eksempel mellom makrell og fotballaget Start. Innputtromma som ledda aktiverer, utdyper ikke fullt ut den strukturen som fins i blandingsrommet, og dermed kan man si at komposisjonen påkaller en såkalt *framvoksende struktur*, altså at deler av betydningsinnholdet ikke kan spores tilbake til ledda.

I den ovennevnte komposisjonelle tilnærminga er det ytterst få sammensetninger som avviker fra regelen om komposisjonalitet. Veletablerte sammensetninger slutter ikke å være komposisjonelle, men prosesseringa av dem blir automatisert i takt med at skjemaene de aktiverer, blir sterkere innprenta hos språkbrukerne. Ved gjentatte eksponeringer for sammensetninga *sykkelpumpe* eller *regnskog* slutter vi å foreta fortolkninger som bygger på en eller annen form for mental sammenpusling av komponentene – for puslespillet er simpelthen lagt. Gjennom forutgående innprenting har sammensetninga blitt en *betegnelse* (jf. 2.4 og Schmid (2020, 262–264)), og fortolkes dermed på lik linje med simplekser

[10]Empiri som sannsynliggjør denne analysen, er at sykkelslanger nesten konsekvent refereres til med uttrykk som konvensjonelt viser til noe som er begrepslig og rent fysisk nærliggende, men ikke eksakt det samme, i utsagn som *å ha flatt dekk* eller *pumpe luft i hjula*. I begge tilfellene er det strengt tatt slangen som er flat eller pumpes opp.

(se blant andre Bakken (1998, 61) og Eik (2019, 226) for samme poeng om maksimalt etablerte sammensetninger). Det vil med andre ord si at de aktiverer ett enkelt innputtrom, ett domene, én kunnskapsstruktur. Sammensetninga har smelta sammen til én nøkkel i Langackers termer. Men dette betyr ikke at vi i etterhånd er forhindra fra å gjøre bedømmelser av gjennomsiktigheten i sammensetninger (Svanlund, 2002), det vil si bedømme i hvilken grad ledda hjelper oss å komme fram til en av de etablerte tolkningene av uttrykket (se videre diskusjon i neste underkapittel). For å unngå å gjøre en regel–liste-feilslutning (Langacker, 1987, 29) der etablerte sammensetninger står på en mental liste, mens uetablerte sammensetninger pusles sammen via skjematiske regler, kan vi si det slik at det er en positiv korrelasjon mellom hvor innprenta en sammensetning er, og i hvilken grad den tolkes via vårt språklige minne knytta til sammensetninga som helhet.

2.5.3 Gjennomsiktighet

Der termen *komposisjonalitet* som nevnt aktiverer en byggeklossmetafor, aktiverer *semantisk gjennomsiktighet* en visuell metafor der språklige uttrykk forstås som fysiske gjenstander eller beholdere med et mer eller mindre anskuelig innhold. Begge termer brukes i forbindelse med komplekse uttrykk bestående av flere betydningskomponenter. Et gjennomsiktig uttrykk, for eksempel en sammensetning, har grovt sett en denotasjon som er i tråd med eller som på et eller annet intuitivt vis følger av denotasjonen til betydningskomponentene.*Komposisjonalitet* og *gjennomsiktighet* er helt klart positivt korrelerte, men det er samtidig fullt mulig å tenke seg at en sammensetning kan være komposisjonell og samtidig ugjennomsiktig, for eksempel hvis komposisjonen bygger på særlig utilgjengelige betydninger av de inngående ledda, som for eksempel betydninga 'framstille pessimistisk' av *svartmale* (se BOB).

Semantisk gjennomsiktighet blir av Downing (1977) (som av Bauer (1983) og Bakken (1998)) satt i sammenheng med etablering (de refererte verka bruker termen *leksikalisering*), der de mest etablerte sammensetningene har en fast forbindelse mellom form og innhold som likner på den ugjennomsiktige og arbitrære forbindelsen mellom et simpleks og dets betydning. Downing foretar et tredelt eksperiment bestående av en benevnelsesoppgave (naming task), en tolkningsoppgave og en rangeringsoppgave. Funna fra det tredelte eksperimentet illustrerer at også nylagde, uetablerte sammensetninger kan ha varierende grad av gjennomsiktighet, og Downing formulerer tre kriterier (eller innskrenkninger) for at en nylaging (novel compound) skal være gjennomsiktig nok til å fungere som et adhoc-ord med klare fordeler i forhold til et nyoppfunnet simpleks eller et konvensjonelt, men upresist simpleks: 1) at sammensetninga utnytter informasjonsressurser, 2) at den er mulig å tolke, og 3) at den denoterer en relevant kategori. Det første

punktet dreier seg om at sammensetninga må angi en mer spesifikk denotasjon enn det det semantiske hodet gjør aleine. Downing bruker sammensetninga *vindflagg* (*wind-flag*) som et eksempel der modifikatoren *vind* ikke gir noen nærmere spesifisering av klassen flagg, siden «alle flagg er ute i vinden» (Downing, 1977, 29).[11] Punkt 2) dreier seg om at sammensetninga må være basert på en relasjon som inngår i kunnskapsbasen til den aktuelle mottakeren. Desto mer kontekstuell og mindre allmenngyldig denne relasjonen er, desto snevrere er det kontekstuelle vinduet der sammmensetninga kan bli korrekt fortolka av en mottaker. Sammensetninga *apple hat* 'eplehatt' er for eksempel tolkbar i flere kontekster om den refererer til 'en hatt med epler på' eller 'en hatt som er forma som et eple' enn om den refererer til 'en hatt det på et gitt tidspunkt fantes et eple under' (til forskjell fra for eksempel *pærehatten* som i samme kontekst skjulte en pære). Punkt 3) knytter seg til hvorvidt sammensetninga benevner en unik kategori med en stabil eksistens. Om kategorien ikke allerede har en konvensjonell benevnelse, har den i det minste et visst potensial for å få det. I eksempelet med eplehatten er det igjen lettere å tenke seg *eplehatt* som en konvensjonell benevnelse for hatter med bilde av epler på enn hatter med epler under fordi det førstnevnte har en mer stabil eksistens.

Libben (1998, 32) hevder på sin side at alle nylagde sammensetninger er gjennomsiktige, men at betydninga til nylagde former over tid forskyver seg bort fra betydninga til sammensetningskomponentene. Dette perspektivet på gjennomsiktighet er så å si identisk med ideen om at grad av semantisk gjennomsiktighet er et produkt av det aktuelle ordets etableringssgrad (jf. blant annet Bakken (1998); Bauer (1983); Downing (1977)). Videre framholder Libben (1998) at sammensetninger kan være ugjennomsiktige, eller opake, i ett eller to sammensetningsledd. Sammensetninga *strawberry* 'jordbær' har for eksempel et relativt opakt forledd og et relativt gjennomsiktig etterledd, mens sammensetninga *shoehorn* 'skohorn' har et mer opakt etterledd. Sammensetninger kan også være relativt ugjennomsiktige, eller ikke-komposisjonelle som helhet. For eksempel er *bighorn* 'tykkhornsau' en forholdsvis ugjennomsiktig sammensetning siden den ikke denoterer et stort horn.

En innvending mot at etablerte sammensetninger har simplekse tolkninger, er at visse studier (blant annet Jarema, Busson, Nikolova, Tsapkini og Libben (1999) og Libben et al. (2003)) finner at priming avhjelper prosesseringa av både gjennomsiktige og ugjennomsiktige (og presumptivt leksikaliserte) ledd. I Libben et al. (2003) forkortes reaksjonstidene i en leksikalsk gjenkjennelsesoppgave (lexical decision task) når respondentene blir eksponert for en priming (altså en forutgående stimulus) av for- eller etterledd, uavhengig av hva slags sammensetning det er. På den annen side finner Sandra (1990) at primingeffekten kun opptrer ved gjennomsiktige sammensetninger. Sistnevnte virker

[11] Sitatet må ikke leses for bokstavelig. Alle flagg er så klart ikke ute i vinden, men *vind* er ifølge Downing ubrukelig til å disambiguere ulike slags flagg.

å ha et klart mer valid eksperimentdesign om man skal få svar på hva slags semantiske aktiveringer som skjer i møte med sammensatte ord av ulik etableringsgrad. Funna til Sandra indikerer at eventuelle assosiative koplinger til for- eller etterledd i ugjennomsiktige (og presumptivt etablerte) sammensetninger ikke er umiddelbare eller sterke nok til at det gir tydelige reaksjonstidsvariasjoner (i alle fall ikke innenfor et eksperimentelt design der man regner i millisekunder). Dette styrker hypotesen om at etablerte komplekser i alle fall umiddelbart har simplekse tolkninger. Men dette betyr slett ikke at den komposisjonelle strukturen ikke er tilgjengelig – med litt betenkningstid. For eksempel ironiserer den britiske komikeren Michael Mcintyre (2019) over deskriptiviteten til sammensatte amerikansk-engelske ord som *sidewalk* (fortau) (på britisk-engelsk *pavement*), *eye glasses* (briller) (på britisk-engelsk *glasses*), *waste paper basket* (papirkurv) (på britisk-engelsk *bin*) og *racket ball* (squash) (på britisk-engelsk også *squash*). Hele det humoristiske poenget til Mcintyre her er at de ekvivalente britisk-engelske orda er simplekser, eller i det minste avledninger, mens amerikanerne – i sin angivelige enfoldighet – behøver mer informasjon for til dømes å fatte hvor et fortau (*sidewalk*) befinner seg, eller hvor brillene (*eye glasses*) skal bæres. Her er det soleklart at Mcintyre gjør komplekse tolkninger av etablerte amerikansk-engelske sammensetninger. Eller for å si det med Bundgaard et al. (2006) og Fauconnier og Turner (2008): Konstruksjonen aktiverer hos Mcintyre (og hans tilhørere) et mentalt skjema bestående av minst to innputtrom.

Borque (2014, 254) formulerer følgende definisjon av semantisk gjennomsiktighet:

> For a lexical unit C, semantic transparency refers to the degree of semantic interpretability of C.

Gjennomsiktighet blir her sidestilt med tolkbarhet (til forskjell fra Langacker (1987) som knytter tolkbarhet til komposisjonalitet, se underkapittel 2.5.2). Borque spesifiserer videre fire premisser for denne definisjonen. 1) Forholdet mellom en sammensetning og dens ledd kan være ikke-eksisterende eller ufullkomment. Selv om det fins en relasjon mellom helhetsbetydninga og sammensetningsledda, er det sjelden slik at denne relasjonen lykkes i å kommunisere alle egenskaper ved sammensetningas betydning eksplisitt, det fins alltid en diskrepans mellom uttrykk og innhold. I ekstreme tilfeller kan sammensetningsledd ha minimal eller ingen påvirkning på sammensetningas betydning, slik som med *løvetann* (mitt eksempel). Dette betyr likevel ikke at sammensetninger ikke kan være gjennomsiktige, bare at gjennomsiktighet er et relativt fenomen. 2) Opplevelsen av semantisk gjennomsiktighet vil uvegerlig variere mellom ulike språkbrukere. I kontekst av en teoretisk modell for hva gjennomsiktighet er, må man imidlertid abstrahere bort denne variasjonen. I denne konteksten er det nødvendig å ta utgangspunkt i en definisjon av gjennomsiktighet der man utgår fra språkbrukere som kjenner betydninga til sammensetningsledda. 3) Semantisk gjennomsiktighet angår både nylagde og etablerte

sammensetninger, men ikke på nøyaktig samme måte. Siden nylagde sammensetninger ikke har en etablert betydning, blir gjennomsiktighet i så henseende et spørsmål om hvor sannsynlig det er at en nylagd sammensetning tolkes i tråd med avsenderens intensjon, gitt betydninga til ledda. For etablerte sammensetninger er det snarere et spørsmål om betydninga til sammensetninga kan utledes fra ledda. 4) En sammensetnings gjennomsiktighetsgrad er i stor grad et spørsmål om mottakers evne til å komme fram til «riktig» betydning basert på sammensetningas komponenter og den aktuelle konteksten. Den mest gjennomsiktige sammensetninga er den som ikke behøver noen spesifikk kontekst for å bli tolka «rett». De fire punkta til Borque (2014) definerer i stor grad gjennomsiktighet som en opplevelse hos og på tvers av språkbrukere og språklige kontekster.

2.5.4 Motivasjon

Samtidig som det er noe omstridt hva som ligger i begrepa KOMPOSISJONALITET og GJENNOMSIKTIGHET, er *motivasjon* en mindre omdiskutert term som også betegner forholdet mellom betydninga til en sammensetning og dens ledd. I Borques punkt 3) fra forrige avsnitt, «semantisk gjennomsiktighet angår både nylagde og etablerte sammensetninger, men på forskjellige måter», overser han at sannsynligheten for at en nylagd sammensetning tolkes i tråd med avsenderens intensjon, trolig påvirkes av den samme faktoren som gjør en etablert betydning mulig å utlede fra ledda. Ifølge Svanlund (2002) er denne faktoren *motivasjonsgrad*. Dette begrunner Svanlund med at verken etablerte, nylagde, gjennomsiktige eller komposisjonelle sammensetninger har betydninger som er forutsigbare. Når for eksempel Bakken (1998, 72) argumenterer for at *damesko* er en komposisjonell og forutsigbar sammensetning, tar hun ifølge Svanlund utgangspunkt i en allerede etablert betydning for damesko, uten å utforske hva slags mulige betydninger ordet kan tenkes å ha. Svanlund trekker for eksempel fram at damesko gjerne står i motsetning til en annen type sko (herresko, pikesko, hverdagslige sko for damer). Uten en spesifikk kulturell erfaring er det heller ikke godt å si hva som gjør en sko «damete», altså er det ikke åpenbart hva slags ekstensjon *damesko* isolert sett har. Konklusjonen til Svanlund (2002, 24) er at betydninga til normale sammensetninger ikke kan sies å være forutsigbar med utgangspunkt i de inngående ledda, men at den er delvis sporbar (partiellt härledbar) til sammensetningsleddas konvensjonelle betydninger, altså at helheten framstår relativt *motivert* av delene. Motivasjon dreier seg her helt enkelt om i hvilken grad det fins en identifiserbar sammenheng mellom sammensetningas og sammensetningsleddas mer eller mindre etablerte betydninger. Dersom det er en sterk sammenheng mellom leddas konvensjonelle betydninger og sammensetningas helhetsbetydning, kan vi si at ledda i stor grad hjelper oss å komme fram til rett tolkning av sammensetninga. Altså er valg av sammensetningsledd relativt motivert gitt sammensetningas intenderte

betydning. Bakkens eksempel *damesko* har forholdsvis motiverte ledd. Man kan for eksempel tenke seg at en skoprodusent kaller et skopar for damesko av den enkle grunn at skoen er produsert for å passe til voksne kvinner, og at det er vesentlig at kunden oppfatter nettopp denne kvaliteten ved skoen. Valget av ledda *dame* og *sko* er derfor relativt motivert. Dette er imidlertid ikke det samme som at *damesko* er en fullt ut gjennomsiktig sammensetning. En kunde kan helt fint ta betegnelsen til inntekt for at skoen først og fremst er mynta på voksne kvinner av fornem herkomst, jf. betydning 2 av *dame* i BOB.

En annen ting som gjør at motivasjon ikke automatisk leder til gjennomsiktighet, er at sammensetninga kan aktivere to domener som vanskelig lar seg forene. Sammensetninga *glassflaske* (mitt eksempel) er muligens selvforklarende for språkbrukeren siden forleddet *glass* spesifiserer et nærliggende felt innenfor flaske-domenet, nemlig flaskens materiale. Helt åpenbart er det likevel ikke, for forleddet kan også spesifisere andre felt, for eksempel flaskens innhold eller funksjon, slik som vi finner i sammensetningene *brusflaske* og *varmeflaske*. En annen teoretisk mulighet er at forleddet spesifiserer for eksempel en likhet, som i *sjirafflaske* (mine eksempler).[12] Om man bruker *sjirafflaske* for å betegne en flaske som likner på en sjiraff, for eksempel ved å ha lang hals, er det ingenting umotivert ved valget av sammensetningsledd. Likevel virker det sannsynlig at *sjirafflaske* vil oppfattes som mindre gjennomsiktig enn for eksempel *glassflaske*, da førstnevnte sammenslutning av domener er mindre umiddelbar og i mindre grad knytter seg til skjematiske mønstre man finner i andre sammensetninger med flaske som etterledd.

Motivasjon er en gunstig variabel fordi det ikke betegner en opplevelse av tolkbarhet, slik som gjennomsiktighet, men snarere betegner en identifiserbar relasjon mellom ledd og sammensetning. Motivasjon er likevel med på å predikere hva den opplevde gjennomsiktigheten hypotetisk vil være for en gjennomsnittlig språkbruker, i og med at motivasjonen estimerer i hvilken grad ledda bistår språkbrukeren i å identifisere «rett» tolkning. En problemstilling knytta til motivasjon er imidlertid hvordan man håndterer polyseme ledd. I visse tilfeller kan et ledd bidra med ulikt betydningsinnhold i ulike sammensetninger, uten at noen av sammensetningene har redusert motivasjonsgrad. Antakelig er både *kongekrone* og *femkrone* relativt motiverte sammensetninger, selv om etterleddet *krone* bidrar med ulike konvensjonelle betydninger i de to tilfellene. Videre kan konteksten også spille inn på hva slags leddbetydning som framstår motivert. Den politiske betydninga av *blå* er trolig mindre tilgjengelig enn fargebetydninga uten en spesifikk kontekst, men om man snakker om en *blåblå regjering*, er kanskje den politiske betydninga til og med mer motivert enn fargebetydninga (se dessuten liknende diskusjon knytta til delstudie 1 i delkapittel 4.1).

[12] *Sjirafflaske* brukes i en artikkel av Stavanger Aftenblad der sosiolog Trond Blindheim kritiserer en trend med kjendisviner som han mener minner om «likører på sjirafflasker» (Frøyland, 2012).

Uansett hvilken av de ovennevnte innfallsvinklene til å beskrive semantisk anomalisering en velger, gir det mening å plassere sammensetninger langs et kontinuum basert på hvor tolkbare deres betydninger er gitt ledda de består av, altså hvordan de oppleves av en gjennomsnittlig språkbruker. Et populært standpunkt er at de minst tolkbare sammensetningene har simplekse tolkninger på bakgrunn av høy etableringsgrad (se f.eks. Bakken (1998); Bauer (1983); Downing (1977); Eik (2019); Sandra (1990) og til en viss grad Bundgaard et al. (2006); Langacker (1987)). Bundgaard et al. (2006); Langacker (1987); Nesset (2017) vil på sin side hevde at de aller fleste sammensetninger er komposisjonelle, men at noen har lavere analyserbarhet (Langacker) eller påkaller en framvoksende struktur (Bundgaard og Nesset). Til sist vil Svanlund (2002) hevde at det varierer hvor innlysende sammenhengen mellom leddas og sammensetningas betydninger er. Siden sammensetninger per definisjon har komplekse morfologiske strukturer, vil jeg betrakte sammensetninger med simpleks semantikk, lav analyserbarhet eller lav motivasjonsgrad som anomalier.

2.5.5 Formell anomalisering

Sammensetninger kan være anomalier med hensyn til formelle trekk. Eik (2019, 225) påpeker at distinksjonen mellom gjennomsiktige og ugjennomsiktige sammensetninger stundom avtegner seg i formelle morfologiske trekk. For eksempel bøyes ifølge Eik sammensetninga *løvetann* ulikt etterleddet sitt, altså uten vokalendring i flertall *løvetanner*, *løvetannene*. Men om man legger en gjennomsiktig betydning til grunn, altså *løvetann* i betydninga 'tann hos løve', vil flertallsbøyninga følge samme mønster som etterleddet *-tann*. Løvetann-eksempelet fungerer om man ser til bokmålsnorma per 2019, men Eik overser med dette eksempelet at nynorsknorma har samme paradigme for *tann* og *løvetann*, altså med vokalendring i siste ledd. For øvrig vedtok Språkrådet i juni 2023 å endre bokmålsnorma slik at *løvetenner* blir normert også for den botaniske betydninga (Språkrådet 2023). Begrunnelsen for dette er blant annet at flertallsforma *løvetenner* forekommer i bokmålstekster i begge betydninger. Poenget til Eik står seg like fullt; bedre eksempler på formelle anomalier er sammensetningene *overvære*; som vanskelig kan bøyes på linje med etterleddet i presens (det skulle blitt **overer*), *brennevin*; som til forskjell fra etterleddet er i nøytrum og *barsel*; som er en redusert form av *barnsøl*. Uavhengig av den bakenforliggende årsaken er alle disse anomalier på formnivå, hvilket formodentlig gjør dem til ganske selvskrevne ordbokoppføringer.

Fonologiske endringer kan også forekomme. Om man sammenlikner sammensetningene *jordbær* og *jordhaug*, så uttales disse av mange med tonemforskjell og ulik vokallengde i forleddet, der førstnevnte har kort vokal til forskjell fra konvensjonell uttale av *jord*.[13]

[13]Men fonologiske endringer er på ingen måte en uunngåelig følge av etablering eller ugjennomsik-

Uavhengig av hva som driver formell anomalisering, blir resultatet en form som det er all mulig grunn til å tro at ordbokbrukere kan ha nytte av å finne informasjon om i ordbøker. Tilsvarende kan man si at sammensetninger med høy semantisk anomaliseringsgrad prinsipielt sett har samme status, selv om det i dette tilfellet er mer diskutabelt hvordan man identifiserer semantisk anomali. Uansett er det en slags konsensus i sammensetningslitteraturen om at ikke alle sammensetninger er likestilte med hensyn til anomaliseringsgrad, og derfor er dette en nyttig variabel for leksikografer som forsøker å identifisere hensiktsmessige sammensetninger å innlemme i ordbøker.

2.6 Utbredelse

På samme måte som semantisk og formell anomalisering indikerer gjennomsiktighet, så kan utbredelse indikere etableringsgrad. Enhver bruksbasert tilnærming til lingvistisk undersøkelse er på et eller annet nivå nødt til å forholde seg til grad av utbredelse for lingvistiske fenomen. Om man ønsker å si noe om hvor etablert, innprenta og konvensjonelt et fenomen er sammenlikna med andre, må man på et eller annet vis forholde seg til hvor ofte disse fenomenene forekommer. Hvis man for eksempel vil undersøke den geografiske spredninga til apikal l (populært kalt *Østfold-l*) i koda, må man uunngåelig ta i betraktning hvor hyppig denne l-en forekommer i denne posisjonen på ulike geografiske plasser. Å måle tilstedeværelse og utbredelse av lingvistiske fenomen i en språklig varietet kan sammenliknes med å måle laksebestanden i fjorden eller revebestanden i skauen, man må uvegerlig telle. Spørsmålet er da hvor mange lakseindivid man finner innenfor et visst avgrensa område. Om man finner 0,3 laks per kubikkmeter vann i en representativ del av fjorden, har man et estimat over tettheten på lakseindivid i fjorden, uten at man trenger å vite hvor stor fjorden er, eller hvor mange lakseindivid det er totalt.

I det følgende vil jeg komme inn på noen helt sentrale problemstillinger knytta til å måle den overordna hyppigheten til lingvistiske fenomen i usus. Variabelen *utbredelse* benyttes her som paraplyterm for ulike måter å beregne et fenomens avtrykk i korpus eller usus på. De første avsnitta nedenfor knytter seg til korpusbruk uavhengig av hvilken metode man benytter. Deretter går jeg mer i detalj på to ulike parametre for utbredelse, nemlig korpusfrekvens og -spredning.

tighet. En sammensetning som *jordmor* må kunne sies å være relativt ugjennomsiktig, men det er helt normalt å uttale denne med lang vokal i første og andre ledd. Videre kan man innvende at sammensetninger som *mormor* og *farmor* derimot regelmessig blir uttalt med kort vokal i forleddet, og at forleddet i disse har mer med sammensetningas betydning å gjøre enn forleddet i *jordmor*. Sammensetningene *mormor* og *farmor* er såleis kanskje mer motiverte og derfor mer gjennomsiktige enn *jordmor*.

2.6.1 Vi har ikke tilgang på «hele språket»

Språkbruken er i statistiske termer en populasjon. Siden vi ikke har tilgang på hele populasjonen, må vi nøye oss med å undersøke hva utvalg av denne populasjonen kan fortelle oss om helheten. I moderne leksikografi, og i bruksbasert lingvistisk metode, er det ofte et eller flere tekstkorpus som tjener som utvalg. Her diskuteres forholdet mellom korpus og usus.

Det ligger til korpusets natur, til forskjell fra språklige varieteter, at det er finitt og derfor mangelfullt (se f.eks. Stefanowitsch (2020, 5-6)). Men korpus kan i større eller mindre grad være representative for den språklige varieteten som korpusmaterialet er samla fra. Representativitet er som Leech (2007) formulerer det, «the holy grail of corpus linguistics». Et representativt korpus gir innlysende nok bedre grunnlag for å trekke slutninger om populasjonen enn et skeivt og mangelfullt ett. Samtidig kan selvfølgelig ulike korpus ha til hensikt å representere ulike populasjoner.

Taylor (2012, 13-15) spør seg hva korpuset dypest sett bør representere. Om det er den lingvistiske erfaringa til språksamfunnets individer, fordrer det korpus med en representativ sammensetning av registre. Her må man for det første få til en hensiktsmessig fordeling mellom skriftlig og muntlig språkbruk (noe de færreste korpus gjør), og for det andre sørge for en hensiktsmessig fordeling av skriftlige og muntlige registre. Et problem er at vi ikke veit andelen av skriftlig og muntlig språkbruk en gjennomsnittlig språkbruker blir eksponert for (en fordeling som trolig vil variere mellom individ avhengig av for eksempel hva slags yrke de har, og hvor gamle de er) (Stefanowitsch, 2020, 29), eller hvilke skriftlige og muntlige registre de oftest eksponeres for og produserer innenfor. Siden det er all grunn til å tro at det fins stor variasjon i den lingvistiske erfaringa til ulike individer innenfor et språkmiljø, får man som Taylor påpeker (2012, 15), et behov for å ivareta representativitet på to nivåer om man skal forsøke å kompilere et fullkomment representativt korpus: Man må samle et representativt utvalg språklige stimuli fra et representativt utvalg individer.

Representativiteten kan videre knyttes til minst to nivåer av korpusets tekstmateriale, nemlig *type* og *eksemplar*. LBK består for eksempel av et balansert utvalg av bokmålstekster med hensyn til tekstkategori. Korpusmaterialet i LBK er med andre ord ment å ha representativitet på typenivå; det inneholder et representativt utvalg type tekster basert på en spørreundersøkelse om lesevaner (Fjeld et al., 2020). Det er prinsipielt umulig at tekstutvalget består av representative eksemplarer av sine typer. Det finnes for eksempel ingen faglitterær eller skjønnlitterær tekst som aleine kan gi en fullgod representasjon av fag- eller skjønnlitteratur. Selv om man kan etterstrebe balanserte fordelinger i henhold til en del variabler som for eksempel tid, kjønn, emne, osv., må man på et eller annet tidspunkt velge for eksempel hvilke stykker av Ibsen og Bjørnson som skal med.

Valget av eksemplarer vil til syvende og sist ha en tilfeldig påvirkning på utbredelsen til ulike n-gram i korpuset. Paulsen (2022) eksemplifiserer dette poenget med å påpeke at hvorvidt korpuset inneholder en biografi om sjøfareren Willem Barentsz eller ei, i det minste har noe å si for hvor utbredte n-gramma *Willem* og *Barentsz* er i korpuset.

I tillegg baserer LBK seg på lesevaner og ikke skrivevaner, altså forsøker det å være representativt i henhold til den språkbruken folk eksponeres for. Dette er trolig noe ganske annet enn å fange den språkbruken en gjennomsnittlig språkbruker produserer. For eksempel utgjør trolig aviser og bøker en vesentlig større del av den skriftlige resepsjonen enn den skriftlige produksjonen til gjennomsnittlige språkbrukere. Sagt på en annen måte leser de fleste av oss trolig flere bøker enn vi skriver.

En av mange kjensgjerninger som kan utledes fra de ovennevnte representativitetsproblemene, er at ususutbredelse på ingen måte er noen direkte refleksjon av korpusutbredelsen. Der usus rommer hvor ofte i språkvarietetens eksistens for eksempel et ord har blitt brukt med en gitt betydning, innenfor en gitt kontekst eller i en bestemt grammatisk konstruksjon, er korpusutbredelse et mål på hvor ofte noe forekommer innenfor en nøye utvalgt, men likevel skeiv og mangelfull tekstmengde, som for øvrig er mikroskopisk i størrelse sammenlikna med usus.

2.6.2 Hvordan identifisere relevante forekomster i korpus?

Det er ikke bestandig helt trivielt å avgrense for eksempel hvilke former som inngår i det fenomenet man ønsker å telle forekomstene til. Når man måler korpusutbredelsen til et bestemt leksem, har det mye å si hva som regnes som forekomster av dette leksemet. Et spørsmål som ikke diskuteres noe særlig innenfor korpuslingvistikken, er kort sagt hvordan man identifiserer forekomster av for eksempel et ord. Hva som regnes som forekomst, er i stor grad påvirka av hvordan man håndterer variabilitet knytta til bøyning, avledning, sammensetning, homonymi, polysemi og diakroni. Hvor inkluderende eller ekskluderende man er overfor slik variabilitet, avgjøres i hovedsak av hva formålet med utbredelsesberegninga er. Om man beregner utbredelsen til et leksem, er det vesentlig at man inkluderer alle bøyningsvarianter av dette leksemet.

For substantiv er det som oftest ukontroversielt hvilke ordformer som hører til leksemet, men for et verb som *løpe* er det ikke selvsagt hva man skal gjøre med for eksempel partisippformen *løpende*. Det virker greit at konstruksjoner som «komme løpende» bidrar til utbredelsen til leksemet *løpe*, men mindre greit med konstruksjoner som «på løpende bånd». Her må man avgjøre om sistnevnte bruk av *løpende* er tilfelle av polysemi og derfor innenfor løpe-leksemet, eller om det er tilfelle av homonymi der *løpende* er et adjektiv som står utenfor verb-leksemet *løpe* (se for øvrig Kinn (2014) for detaljert diskusjon om verb- og adjektivegenskapene til presens partisipp).

Avledning skaper hodebry for tellinga blant annet fordi veldig mange avledninger er systematisk relatert til sitt grunnord. Laufer og Nation (1995) opererer for eksempel med en samlekategori for grunnord og deres eventuelle avledninger kalt «word families». De fleste verb har et korresponderende verbalsubstantiv, slik som *samle* har *samling*, men visse verbalsubstantiv virker mer selvsagte som egne leksem enn andre. For eksempel er *samling* i betydning 'noe som er samlet' (se BOB) et mer selvstendig leksem enn ordformen *taing* til verbet *ta*. Videre kan veldig mange adjektiver, som f.eks. *slurvete*, brukes adverbielt, men det er usikkert om den adverbielle bruken gir grunnlag for å si at man har to separate leksem, altså et adjektivleksem og et adverbleksem.

På samme måte som med avledninger kan det i visse kontekster være rimelig at sammensatte orddanninger spiller inn på hva en regner som korpusutbredelsen til en mindre ordstamme, for eksempel at *ertesuppe* og *blomkålsuppe* spiller inn på utbredelsen til *suppe*. De sammensatte ordformene har i det minste en relevans dersom man vurderer ordbokkandidaturet eller innprentingsgraden til *suppe*. I visse tilfeller kan imidlertid sammensetninga ha høyere utbredelse enn sine komponenter, for eksempel er *brøkdel* mer utbredt enn *brøk* i LBK. Visse komponenter, f.eks. *ut* eller *bak*, kan være så produktive i sammensetninger at det blir svært upraktisk å beregne utbredelse basert på alle deres videre orddannelser.

Allomorfi skaper også hodebry for utbredelsesberegninger. For diakrone studier kan det være langt fra trivielt å slå fast hvilke former som tilhører hvilket leksem på et seinere språkstadium. Synkront fins det dessuten også store innslag av allomorfi i ordforrådet, knytta til f.eks. rettskrivning og dialektbruk. Et ekstremt eksempel er sammensetninga *lavtlønnet*, som har 12 ulike normerte skrivemåter i bokmål (se BOB).

Homonymi og polysemi er beslekta problemstillinger i utbredelsesberegninger. Om to svært ulike betydninger kan uttrykkes med samme form, er det ikke alltid helt trivielt å slå fast om vi har med to homonyme ord å gjøre, eller om det skal behandles som ett ord med forskjellige betydninger.

Spørsmåla i avsnitta ovenfor besvares nok helst ut fra hva slags forskningsspørsmål man har. Det er likevel slik at tellinga av forekomster kan operasjonaliseres ulikt i ulike sammenhenger.

2.6.3 Korpusfrekvens

Som Gries (2008) påpeker, er korpusfrekvens den klart mest brukte målemetoden innenfor korpuslingvistiske tilnærminger. I det følgende diskuteres to sider ved frekvens som statistisk mål på hyppighet i usus, heretter kalt *ususfrekvens*.

Operasjonalisering av frekvens

Som nevnt ovenfor i underkapittel 2.6.1, er korpus et statistisk utvalg som benyttes til estimere fordelinger i populasjonen. Korpusfrekvens er såleis et mål på ususfrekvens. Hvor godt korpusfrekvens gjenspeiler ususfrekvens for et gitt n-gram, avhenger delvis av hvordan man normaliserer frekvenstalla slik at de er relative til tekstmengden de er beregna ut fra.[14] Wallis og Mehl (2022, 101) konstaterer at *per ord* og *per millioner ord* er hyppig brukte normaliseringsteknikker innenfor korpuslingvistikk. Dette er trolig både den enkleste og mest intuitive måten å operasjonalisere frekvens på, at divisoren, altså det tallet som antall forekomster deles på, på en eller annen måte tilsvarer korpusets totale størrelse i ord. Mehl (2016); Wallis og Mehl (2022) påpeker imidlertid at denne divisoren har visse svakheter. I probabilitetsteori er divisoren det totale antallet mulige utfall for et fenomen. Et intuitivt tilfelle er antallet ganger man kaster en mynt. I et myntkastscenario korresponderer det teoretisk maksimale antallet forekomster av utfalla krone eller mynt med antallet myntkast. Frekvensen (eller sannsynligheten) av mynt i et gitt eksperiment er derfor antall observasjoner med mynt delt på antall kast. Om vi opererer med feil antall kast, blir nødvendigvis den rapporterte frekvensen upresis og tilnærmet verdiløs. Som Mehl (ibid. 35) skriver:

> If a baseline is inappropriate or incorrect (or absent), then any statistical conclusions drawn from the quantitative data are likely to be unsound.

Et betimelig spørsmål blir da hva som er en fornuftig divisor når man måler korpusfrekvens. Der mynt er et mulig utfall for ethvert myntkast, er ikke ordet *mynt* en mulig forekomst på alle posisjoner i et korpus bestående av autentisk språkbruk. Hvor mange ganger *mynt* i teorien kan forekomme i korpuset, avhenger blant annet av den gjennomsnittlige lengden på frasene i korpuset. Et spørsmål er da om det ikke ville være mer presist å operasjonalisere korpusfrekvensen til *mynt* ved hjelp det totale antallet substantiver, eller det totale antallet nominaler i korpuset. Selv om det selvsagt er umulig å tenke seg at alle nominaler i et korpus er *mynt*, er det vesentlig mer plausibelt enn at alle orda i korpuset er *mynt*. En styrke med å benytte en mer nyansert divisor er at frekvensen til ord fra ulike ordklasser blir mer sammenliknbare. Frekvensen til for eksempel verbet *spankulere* blir da antall forekomster dividert på antall verbfraser i korpuset, hvilket gjenspeiler hvor frekvent ordet er – til verb å være!

En annen fordel med en ordklassebundet frekvensmåling er at man unngår at lister over de mest frekvente orda domineres av for eksempel nominaler, som det i de aller fleste setninger er flere av enn for eksempel verb. I en helt vanlig setning som *Hunden snuste*

[14] *Frekvens* brukes gjennomgående om relative tall i denne avhandlinga. Jeg avstår dermed fra å bruke termene *absolutt frekvens* og *relativ frekvens* (se definisjon i underkapittel 2.7.1.

på hånda til gutten, fins det tre nominale ledd og kun ett verbalt ledd. Om vi tenker oss setninga som et minikorpus, fins det i prinsippet tre mulige brukskontekster for nominaler, to for preposisjoner, og en for verb. Med en ordklassebundet divisor vil altså leksemene *hund*, *hånd* og *gutt* ha lavere frekvens enn *snuse*, siden sistnevnte forekommer i den eneste posisjonen den kan forekomme, mens de førstnevnte forekommer i en av tre mulige kontekster. I prinsippet virker dette rimelig, men man kan spørre seg hvilken praktisk betydning det har. I realiteten er det vel sjelden slik at ett og samme substantiv forekommer på tre av tre mulige plasser i en og samme setning, som i *hunden snuste på hunden til hunden*. På den annen side virker en setning som *Hunden hans snuste på hunden hennes* helt konvensjonell.

Videre kan man innvende mot den ordklassebundne innfallsvinkelen at det er en styrke at korpusfrekvens indikerer utbredelse på tvers av ordklasser i og med at vi for eksempel produserer og eksponeres for flere nominaler enn verb, og at vi ikke har separate konvensjonaliserings- og innprentingsmoduler for ulike ordklasser. Dessuten kan utbredelsen til visse sjeldne ordklasser blåses kraftig opp. Interjeksjoner er for eksempel ganske sjeldne i de fleste tekster. Dette vil nødvendigvis føre til at det totale antallet interjeksjoner er temmelig lavt sammenlikna med for eksempel substantiv eller verb, og at de vanligste interjeksjonene dermed får veldig høy korpusfrekvens sammenlikna med for eksempel middels frekvente substantiv. I slike tilfeller kan sammenlikninger mellom ordklassebaserte frekvenser gi direkte misvisende indikasjoner på hvor ofte en språkbruker produserer eller resiperer visse interjeksjoner versus for eksempel substantiv.

Wallis og Mehl (2022) sitt hovedpoeng er likevel at det ikke er trivielt hvordan man operasjonaliserer korpusfrekvens, og at visse forskningsspørsmål fordrer andre operasjonaliseringer enn de tradisjonelle divisorene som utgjøres av «per–ord» eller «per–million–ord». For eksempel egner andre operasjonaliseringer seg når man skal sammenlikne ulike realiseringer av et avgrensa fenomen, for eksempel hvilke ord som er forbundet med presens partisipp. I slike tilfeller vil det være relevant å ta hensyn til for eksempel hvor mange presens partisipp-former det fins i korpuset, og hvor mange forekomster de aktuelle verbleksemene har totalt sett. Til leksikografisk seleksjon av sammensetninger er man imidlertid interessert i utbredelsen til enkeltord uavhengig av hva slags utbredelse andre ord har. I de fleste tilfeller er man ikke interessert i for eksempel hvor stor andel av substantivforekomstene som utgjøres av *pappskalle*, men bare hva den estimerte ususfrekvensen til *pappskalle* er. Derfor er trolig den tradisjonelle operasjonaliseringa av korpusfrekvens rimelig i møte med den leksikografiske problemstillinga i denne avhandlinga.

Frekvenstall har høy variabilitet

Frekvenstall har en tendens til å variere mye mellom korpus og korpusdeler. I hvilken grad distribusjonen til for eksempel et leksem i et korpus gjenspeiler distribusjonen i usus, er som nevnt ovenfor et spørsmål om korpusets representativitet. Det faktum at fullkommen representativitet på både type- og særlig eksemplarnivå er så godt som umulig å oppnå, gjør at frekvenstall kan variere mye fra korpus til korpus (se f.eks. Gries (2022a); Paulsen (2022)), hvilket gjør frekvensmålinger lite reliable i statistisk forstand. En logisk følgeslutning av denne kjensgjerninga er at ikke alle frekvensmålinger kan være like valide representasjoner av ususfrekvens (se avsnitt om delstudie 2 for inngående diskusjon av dette poenget). Når frekvensmålinger er vidt forskjellige avhengig av datagrunnlaget, kan ikke alle være like treffsikre representasjoner av usus, for ususfrekvensen forblir den samme størrelsen uavhengig av hvilket utvalg man bruker for å representere den.[15]

Frekvenstall er også svært variable på tvers av et korpus' ulike deler (se f.eks. Gries (2022a)). Visse typer vokabular eller konstruksjoner har en klar kopling til visse tekstsjangre, altså ulike typer. Det er selvsagt rimelig at dette gjenspeiles i korpus, men om frekvenstall kun oppgis for korpuset som helhet, altså globalt, kan man gjøre gale antakelser om at det sjangerspesifikke vokabularet er frekvent i språket som helhet.

Dessuten har visse typer vokabular en sterk kopling til ulike enkelttekster, altså eksemplarer. Gries (2022a) finner for eksempel at det engelske ordet *fold* 'brett' noe overraskende er den sjuende mest frekvente imperativformen i hele korpuset ICE-GB, men alle forekomstene kan spores til samme enkelttekst, en bok om origami. De mest frekvente orda i språket har imidlertid liten frekvensvariabilitet mellom korpus og korpusdeler. Dette er gjerne funksjonsord som forekommer mange ganger i de aller fleste tekster, hvilket gjør dem relativt upåvirket av sampling- og sjangereffekter.

En måte å kontrollere for frekvenstalls variabilitet på er å kombinere frekvensmålinger med spredningsmålinger. I det følgende diskuteres spredning som mål på utbredelse i korpus.

2.6.4 Spredning

Spredningsmål er en kategori av statistiske formler hvis utregna verdi gir en indikasjon på n-grams distribusjon på tvers av korpuset. Der frekvens sier noe om hyppighet, sier spredningsmål noe om hvordan hyppigheten er distribuert. Gries (2008) gir en svært grundig oversikt over ulike mål på spredning, og er trolig den som har forska mest på spredning i korpussammenheng i seinere tid (se for eksempel Gries (2008, 2010, 2021, 2022a, 2022b)). Det er likevel verdt å bemerke at for eksempel Juilland, Brodin og Da-

[15] Men ulike forskningsspørsmål kan selvsagt ha ulike deler av usus for øye.

vidovitch (1970) brukte spredning til å utvikle en korpusbasert fransk ordbok allerede på 1970-tallet. Gries (2022a) og Paulsen (2022) sammenlikner forholdet mellom korpusfrekvens og -spredning med forholdet mellom gjennomsnitt og standardavvik i et statistisk utvalg: Førstnevnte indikerer sentraltendens i utvalget, mens sistnevnte indikerer spredning rundt denne sentraltendensen. Spredningsmål kan grovt sagt deles inn i to undertyper, delbaserte mål og distansemål. Førstnevnte, den vanligste typen, måler frekvensvariabilitet på tvers av korpusets metavariabler som f.eks. år, domene, forfatter, dokument eller liknende (se Gries (2021); Juilland et al. (1970)). Sistnevnte ignorerer denne strukturen og måler spredning basert på distansen mellom for eksempel hver gang et leksem forekommer (se Savický og Hlavácová (2002)) dersom man omgjør korpuset til en ustrukturert sekk med ord (bag of words).

Det er gunstig å kombinere frekvensmålinger med spredningsmålinger siden førstnevnte gir informasjon om størrelsesordenen til en distribusjon, mens sistnevnte kontrollerer for om denne størrelsesordenen er et resultat av hyppighet på tvers av metavariabler, eller om den er sterkt overrepresentert innenfor for eksempel en sjanger eller liknende. Spredningsmål kontrollerer såleis også for om frekvensen på et eller annet vis er en arbitrær effekt av utvelgelsen av tekster til korpuset. Dersom en sammensetning forekommer hyppig i kun et par dokumenter, er tilstedeværelsen til disse dokumenta helt avgjørende for korpusfrekvensen til sammensetninga. Men om sammensetninga forekommer i nærmest alle dokumenta i korpuset, er det forholdsvis sannsynlig at man ville funnet omtrent samme korpusfrekvens også med en annen sampling.

En annen grunn til at høy korpusfrekvens og jevn spredning med all sannsynlighet indikerer høy ususfrekvens er at ord som vi veit har høy ususfrekvens gjennomgående har høy korpusfrekvens og jevn spredning. For eksempel veit vi at konjunksjonene *og*, *men*, *for* og *eller* eller pronomen som *jeg*, *han* og *hun* forekommer relativt mange ganger i de aller fleste tekster og derfor har en høy ususfrekvens. I de aller fleste korpus vil man finne at disse orda har både jevn spredning og høy frekvens.

Utbredelse i korpus er altså et multivariat fenomen som bør måles gjennom både frekvens og spredning (se omtale av alternative målemetoder i underkapittel 2.6.6). Selv om frekvens og spredning ofte er korrelert, er det fullt mulig for et leksem å ha høy frekvens og ujevn spredning eller vice versa. Samla gir de altså et forholdvis godt bilde på utbredelse i usus, og særlig ususfrekvensen. I det følgende diskuterer jeg imidlertid hva man kan utlede av ususfrekvens.

2.6.5 Hva er ususfrekvensen reelt sett et mål på?

Det virker intuitivt at leksikografer har nytte av å vite hvor hyppig f.eks. et ordbokaktuelt ord forekommer i autentisk språkbruk. Desto hyppigere et ord forekommer i

språkbruk, desto større sjanse er det for at ordbokbrukerne støter på dette ordet i produktiv eller reseptiv sammenheng. Men spørsmålet om hvilke ord språkbrukere støter på, er i stor grad knytta til spørsmålet om hvilke fenomener de støter på. Selv om vi umiddelbart kan slå fast at ethvert språk inneholder funksjonsord som ikke har noen åpenbar referent «ute i den virkelige verden», burde det vel være en slags forbindelse mellom et språksamfunns fenomenverden og med hvilken frekvens ulike innholdsord blir tatt i bruk i nevnte språksamfunn. Som Taylor (2012, 146) påpeker, forekommer orda for hund og katt oftere enn orda for sjiraff og jordsvin i engelsk, av helt åpenbare grunner. Men ideen om at ordfrekvens er en refleksjon av fenomenfrekvens, at det fins en slags frekvensikonisitet mellom usus og virkelighet, kan temmelig lett gjendrives ved for eksempel å vise til synonymi. Selv om frekvensen i Nasjonalbibliotekets n-gramsøk indikerer det, kan det umulig finnes eller bli spist flere hyser enn koljer (da betegnelsene *hyse* og *kolje* viser til samme fiskeart). Usus er påvirket av et utall kontekstuelle, sosiale og sjangermessige variabler, og språkbrukerne har dessuten bevissthet om relative frekvensfordelinger mellom ulike typer vokabular – for eksempel at *hund* er vanligere og mindre markert enn *kjøter* – og onomasiologiske valgmuligheter knytta til alle disse variablene. Hvorvidt den ene eller andre varianten blir realisert i en gitt kontekst, påvirkes i større grad av språkbrukerens subjektive språkvalg enn av det objektive fenomentilfanget som eksisterer i den aktuelle situasjonen. Fenomenverdenens innflytelse på frekvensfordelinger kan gjøre seg gjeldende i visse tilfeller, f.eks. finner Stefanowitsch (2005) at differansen i korpusfrekvens mellom setningene «I live in New York» og «I live in Dayton, Ohio» likner på differansen i folketall mellom de to stedene.[16] Selv om det kan virke besnærende at det skulle være en slik korrespondanse mellom korpusfrekvens og reelt folketall, påviser Taylor (2012, 151) at samme korrespondanse forsvinner om man sammenlikner setningene «He lives in New York» og «She lives in New York» eller «He lives in New York» og «He lived in New York». I førstnevnte eksempel ville man antatt omtrent lik frekvens, mens man i sistnevnte eksempel ville forventa en mer frekvent preteritumsform; det er tross alt langt flere mennesker som har bodd i New York en eller annen gang i livet, enn mennesker som bor der akkurat nå.[17] Likevel er setninga med hankjønnspronomen vesentlig mer frekvent enn sin motpart, og det samme med setninga med verbet i presens. Taylor (2012) konkluderer med at frekvensfordelinger i språket først og fremst gir et bilde på språket, dets bruk, og hva slags statistiske intuisjoner en typisk språkbruker besitter med hensyn til alle språklige fasetter, lyder, ord, konstruksjoner, setninger, tekster, etc. Korpusutbredelse er med andre ord ikke et avbilde av fenomenutbredelse. Vi bør dermed ikke konkludere med at *bartre* nødvendigvis er en bedre ordbokkandidat enn *postkontor* bare fordi det eksisterer langt flere eksemplarer av førstnevnte. Som po-

[16] At det skulle være en stor frekvensforskjell på disse setningene, ble først foreslått av Noam Chomsky.

[17] Taylor sier ingenting her om tidfestinga for de ulike forekomstene av «He lives in New York». Om ytringa har skjedd i fortid, kan den ha blitt ytra av en forhenværende innbygger.

engtert i underkapittel 2.2.2 fins det dessuten fenomen som mennesker regelmessig er i befatning med, men som sjelden blir referert til i kommunikative hendelser (Zenner et al., 2014).

Man kan dessuten tenke seg at det er interessant for leksikografer å identifisere de lavfrekvente orda, siden det er disse brukerne med høyest sannsynlighet er ukjente med. Denne antakelsen bygger på et premiss om at bruk leder til sterkere innprenting, jamfør underkapittel 2.2.2. Hypotesen om at det er en nær forbindelse mellom ususfrekvens og innprenting har blitt styrka av et utall psykolingvistiske eksperiment:

> The last 50 years of psycholinguistic research has demonstrated language processing to be exquisitely sensitive to usage frequency at all levels of language representation: phonology and phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension, grammaticality, sentence production, and syntax. (Ellis, 2002, 35)

Likevel har relevansen til frekvens som mål på innprenting til dels blitt underbygget av statistisk uholdbare tilnærminger, jf. Gries (2022a, 51):

> It is easy to obtain significant correlations between frequency as a predictor and some dependent variable because such a test tests the role of frequency against a null hypothesis that frequency plays no role while controlling for nothing else.

Dersom ususfrekvens målt ved korpusfrekvens ikke er ledsaga av andre mål på innprentingsgrad, er det intet mysterium at det viser seg å være korrelert med for eksempel prosesseringshastighet. Det ville vært underlig om grad av innprenting ikke skulle spille inn på hvor raskt respondenter klarer å gjenkjenne for eksempel et ord eller en grammatisk konstruksjon, og det er slett ikke uventa om ususfrekvens i alle fall er korrelert med grad av innprenting. Men når ususfrekvens er den eneste prediktoren i en studie som er ment å indikere noe om innprentingsgrad, sier ikke den signifikante relasjonen mellom korpusfrekvens og responsvariabelen oss noe om styrkene eller svakhetene til ususfrekvens (eller korpusfrekvens) som mål på innprenting. Og dette er i tilfeller hvor det i det hele tatt fins en positiv korrelasjon mellom disse størrelsene. Svanlund (2009) finner for eksempel kun en svak korrelasjon mellom korpusfrekvens og innprentingsgrad hos sine informanter, og konkluderer med at innprentingsgrad ikke utelukkende er et spørsmål om kvantitet, men også avgjøres av kvalitative variabler som oppmerksomhetsverdi (uppmärksamhetsvärde).

Schmid (2020) hevder som nevnt i delkapittel 2.2 at språket er gjenstand for de to overordna hovedprosessene og -effektene konvensjonalisering og innprenting. I Schmids

system har konvensjonalisering og innprenting et felles skjæringspunkt i språkbruk. At konvensjonalisering og innprenting er både prosesser og effekter, har den forklaring at det fins en tilbakekopling i systemet der konvensjonaliserings- og innprentingsgrad både påvirker hva slags språklig innputt mennesker blir eksponert for, og blir påvirket av den språklige utputten språkbrukere produserer. Bruken av en gitt konstruksjon i en gitt kontekst er informert av konvensjonalisering og innprenting, samtidig som den bidrar til nettopp konvensjonalisering og innprenting.

I Schmids modell kan man muligens koble frekvens til konvensjonalisering, og da mer spesifikt til de to delprosessene som konvensjonalisering består av, nemlig *usualisering* og *diffusjon* (Schmid, 2015, 17-18). Førstnevnte sikter litt forenkla til etablering og videreføring av sambandet mellom et språklig uttrykk eller en kommunikativ handling og den kommunikative effekten som følger. Sistnevnte sikter til samfunnsutbredelsen til en usualisert kommunikativ handling (se også forklaring av termene i underkapittel 2.2.1). Forskjell i diffusjon avtegner seg sannsynligvis i ususfrekvens. Høy ususfrekvens følger av høy diffusjon, mens lav ususfrekvens er tegn på lav diffusjon med uviss usualisering.

Schmid selv knytter imidlertid ususfrekvens til det han kaller *repetisjonsfrekvens* (frequency of repetition), som er en viktig faktor for innprenting. Ususfrekvensen indikerer hvor ofte en gjennomsnittlig språkbruker har blitt eksponert for eller produsert et uttrykk. Desto oftere et uttrykk blir fortolka eller produsert for fortolkning, altså repetert, desto dypere innprenta eller internalisert blir det relevante assosiasjonsmønsteret mellom form og innhold for en aktuell språkbruker. Repetisjonsfrekvensen er altså hvor hyppig et gitt individ har blitt eksponert for eller har produsert et n-gram, mens ususfrekvensen er hvor hyppig det aktuelle n-grammet forekommer blant alle brukere av den aktuelle varieteten.

Det er imidlertid grunn til å tro at ususfrekvensen kun er korrelert med innprenting og konvensjonalisering, nærmere bestemt repetisjonsfrekvens og diffusjon, og at etableringsgraden ikke avgjøres av ususfrekvensen aleine. Med eksempelet om fag- og allmennspråk (i underkapittel 2.2.1) i mente dreier ikke diffusjon seg bare om hyppighet, men om overskridelse av sjangre, kontekster og språklige konstruksjoner. Ususfrekvensen til sammensetninga *besøkelsestid* forteller oss for eksempel ikke at den ifølge LBK kun forekommer i det faste uttrykket «kjenne sin besøkelsestid», og da reelt sett har samme diffusjon som det aktuelle uttrykket. Tilsvarende ususfrekvensen bare noe om hvor hyppig de formene som ligger til et leksem, blir brukt i språket. Her er det høyst usikkert om for eksempel ususfrekvensen til formen *opp* forteller oss noe substansielt om hvor innprenta eller diffundert ordet *opp* er. Her skulle en tro det var relevant å for eksempel differensiere mellom dusinvis av ulike fraseverb som *kaste opp* og *gjøre opp*, eller at også de utallige komplekse ordformer som inneholder *opp*, som *oppnå*, *oppdage*, *oppstyr*, spiller en rolle. Det virker dessuten ikke innlysende at ordformene *små*, *mindre* og *minst* øker

innprentinga og diffunderinga til entall positiv *liten*. Ususfrekvensen innvarsler altså innprenting gjennom repetisjonsfrekvens og konvensjonalisering gjennom diffundering, men med mindre vi gjør mer detaljerte studier av for eksempel hvor velkjent et ord et på tvers av språkbrukere, eller hvor raskt en gjennomsnittlig språkbruker gjenkjenner et ord relativt til andre ord, kan vi ikke vite i hvilken grad ususfrekvensen gjenspeiler innprenting og konvensjonalisering.

I dette underkapittelet har vi sett at ususfrekvens i alle fall i visse tilfeller er korrelert med blant annet fenomenfrekvens, innprenting og konvensjonalisering, men at det ikke kan sies å være noe direkte mål på disse fenomenene. Derfor bør man i leksikografisk sammenheng ikke underslå kvalitative variabler som for eksempel indikerer etablering gjennom usualisering.

2.6.6 Alternative målemetoder

Om korpusutbredelse skal indikere ususfrekvens, er det vesentlig at man bruker et mangfold av korpusmetoder. Som en generell rettesnor bør man som Gries (2022a, 50) påpeker, «increase the resolution on the frequency data». Å måle korpusutbredelse og ingenting annet er analogt med å utgå fra at vi ikke har noen forutgående kunnskap om fenomenet vi måler. Men leksikografer har i veldig mange tilfeller en intuitiv fornemmelse av etableringsgraden til en sammensetning. Deres korpusundersøkelser bør derfor innrettes slik at de tester om fornemmelsen er rett, og hva som eventuelt er opphavet til den. For eksempel kan fornemmelsen stamme fra høy diffusjon, usualisering, repetisjonsfrekvens (se Schmid (2020)) eller oppmerksomhetsverdi (Svanlund, 2009). For å måle disse tingene er det vesentlig at man benytter seg av en multivariat metode som innbefatter reliable og valide korpusmetoder og et utvalg kvalitative variabler. I dette delkapittelet og i avhandlinga for øvrig har jeg fokusert på korpusfrekvens og korpusspredning, men det fins også andre korpusmål «på markedet». Noen av disse listes opp her:[18]

- Egbert og Burch (2022) foreslår at *gjennomsnittlig tekstfrekvens* er et bedre mål på *leksikalsk prevalens*, altså hvor viktig et ord er, enn global korpusfrekvens. Gjennomsnittlig tekstfrekvens beregnes ved å dele antall forekomster i hver tekst på de respektive tekstenes lengde, før man til slutt måler gjennomsnittet av disse tekstfrekvensene over hele korpuset. Styrken med dette målet er at det tar utgangspunkt i korpusets byggeklosser. Tekster er til forskjell fra korpus blant annet autentiske (dvs. de forekommer naturlig uten et bakenforliggende lingvistisk forskningsformål) og selvstendige. Språket, i alle fall det skriftlige, er til syvende og sist organisert i tekster, eller i det minste i ulike diskursenheter atskilt av tid og rom –

[18] Se dessuten Gries (2022a) sin omtale av målene *surprisal*, *entropy* og *contextual diversity*.

ikke i balanserte tekst- eller diskurssamlinger (som korpus er). Derfor hevder Egbert og Burch (2022) at frekvens på tekstnivå gir langt mer presis informasjon om språket enn frekvens på korpusnivå.

- McDonald og Shillcock (2001) foreslår et mål kalt *Contextual distinctiveness* som indikerer hvor mye kontekstinformasjon som kan trekkes ut av n-gram basert på kollokasjonsdata. Ord som kun opptrer innenfor veldig spesifikke kontekster, vil presumptivt være omringa av et mer begrensa sett av kollokater enn ord som er troende til å opptre i omtrent enhver kontekst. Verbet *løpe* kan forekomme i nær sagt hvilken som helst kontekst, og man får derfor ingen veldig sterke forventninger til at visse andre ord skal forekomme i umiddelbar nærhet. Adjektivet *løpsk* har derimot en mer begrensa bruk, og man vil forvente at leksemet *løpe* med ganske høy sannsynlighet forekommer innenfor samme setning eller kontekst. Sistnevnte gir derfor mer kontekstuell informasjon enn førstnevnte. Styrken til *contextual distinctiveness*, sammenlikna med korpusfrekvens, er at den tar hensyn til hvordan ord forekommer i autentisk språkbruk, i lag med andre ord. I tillegg, eller nettopp derfor, virker den som en bedre prediktor for reaksjonstid i visse psykolingvistiske eksperiment (McDonald & Shillcock, 2001, 309 ff.).

Gries (2022b) poengterer at uavhengig av hva slags mål man bruker for å utlede innsikter om et språk basert på et korpus, som i statistiske termer tilsvarer henholdsvis en populasjon og et utvalg, så må man utforske og rapportere den statistiske usikkerheten i funna. Videre er det ikke trivielt hva slags usikkerhetsestimater man bruker, for eksempel er parametriske konfidensintervaller basert på normalfordeling uegna i korpussammenhenger siden korpusdata som frekvens og spredning som regel er fordelt etter Zipfs prinsipp snarere enn normalfordelt (Gries, 2022b).[19] Der en normalfordeling inneholder observasjoner som fordeler seg symmetrisk langs en skala med flest observasjoner på midten, har en zipfisk fordeling en l-liknende kurve der brorparten av observasjonene klynger seg i bunnen av skalaen.

I dette delkapittelet har jeg diskutert variabelen korpusutbredelse som indikasjon på ususfrekvens og etableringsgrad, og hvordan man best måler korpusutbredelse slik at målingene ikke i for stor grad blir påvirket av for eksempel arbitrære valg knytta til sampling. I neste delkapittel samler jeg trådene fra de to første kapitla i denne avhandlinga.

[19] Gries (2022b) hevder man heller bør bruke *bootstrapping*-teknikker for å utforske usikkerheten i korpusfunn.

2.7 Oppsummering

I inneværende kapittel har jeg henvist til sentral morfosemantisk og kvantitativ forskning som sammenlagt utgjør en basis for å utvikle metoder for seleksjon av sammensetninger. For å oppsummere avhandlinga hittil vil jeg her definere de viktigste begrepa i denne avhandlinga, før jeg griper tilbake til problemstillinga i kapittel 1 og peker videre mot de neste kapitla.

2.7.1 Begrepsapparat

Den forskningshistoriske gjennomgangen i delkapittel 2.4 og 2.5 illustrerer tydelig at forskninga på sammensetninger, særlig på de semantiske sidene ved dem, er prega av terminologisk variasjon og overlapp. Under følger derfor en liste som redegjør for de viktigste begrepa i denne avhandlinga.

- **Etablering** betegner i denne avhandlinga noe som andre steder (f.eks. hos Bakken (1998); Downing (1977); Eik (2019)) kalles leksikalisering, altså både en utvikling der en leksikalsk enhet, f.eks. et ord, blir suksessivt innarbeida i et språksamfunn og følgelig dets deltakeres språklige minne, og resultatet av en slik utvikling. Denne prosessen med varierende varighet kan videre brytes ned i to distinkte, men overlappende delprosesser kalt *konvensjonalisering* og *innprenting* (etter Schmid (2020)). Etablering har videre som omtalt i delkapittel 2.4 og i Svanlund (2009, 38) både en semasiologisk og en onomasiologisk side.

- **Konvensjonalisering** betegner her den samfunnsmessige og sosiale siden av etablering, den som fører til at forbindelsen mellom et utrykk og et begrep blir til en konvensjon som språkbrukerne innenfor språksamfunnet anvender og dermed viderefører. Konvensjonalisering kan videre brytes ned i delprosessene *usualisering* og *diffusjon*. Førstnevnte betegner kort sagt opprettelsen av en konvensjon, mens sistnevnte betegner spredninga av den. Ordbøker har konvensjonaliserende kraft fordi de kan bidra til å spre informasjon om konvensjoner overfor språksamfunnet.

- **Innprenting** betegner den psykologiske prosessen der sambandet mellom f.eks. et ord og dets betydning konsolideres i det språklige minnet til en språkbruker. Innprenting samvirker med konvensjonalisering gjennom at konvensjonelle uttrykksmåter innenfor et språksamfunn forutsetter at uttrykksmåten er innprenta i et visst antall språkbrukere innenfor språksamfunnet. Mange aspekter kan bidra til å innprente forbindelsen mellom form og innhold i en språkbruker. Hvor mange ganger språkbrukeren har produsert eller blitt eksponert for ordet, spiller helt klart en

rolle. Likeså kan sambandet innprentes gjennom ekstralingvistiske erfaringer, hvor iøynefallende ordet er, oppmerksomhetsverdi (jf. Svanlund (2009)) eller liknende.

- I tråd med forståelsen til Svanlund (2002) og Borque (2014) betegner **gjennomsiktighet** her språkbrukernes hypotetiske opplevelse av hvor tolkbar en gitt betydning til en sammensetning er. Altså med hvilken letthet en gjennomsnittlig språkbruker uten tidligere eksponering for sammensetninga kan identifisere en gitt betydning med utgangspunkt i sammensetningas overflateform. Hvor forutsigbar er for eksempel betydninga 'lett, ermeløs overdel til sommerbruk for kvinner' (se BOB), gitt overflateformen *solliv*? Siden gjennomsiktighet betegner en gjennomsnittlig opplevelse, er det ikke mulig for leksikografer eller språkforskere å gjøre direkte vurderinger av gjennomsiktighet slik det er definert her. Det er snarere noe man må forsøke å predikere ved hjelp av et sett gjennomsiktighetsfaktorer. Det er et viktig forskningsspørmål (som for øvrig stilles i delstudie 1) hvilke faktorer som påvirker gjennomsiktigheten til sammensatte ord, altså den hypotetiske opplevelsen av tolkbarhet. I denne avhandlinga vil jeg utgå fra at opplevelsen påvirkes av sammensetningas semantiske og formelle anomaliseringsgrad, og at opplevelsen av gjennomsiktighet derfor følger av den akkumulerte anomaliseringsgraden. Gjennomsiktighet kan knyttes til etablering gjennom at utstrakt ugjennomsiktighet forutsetter usualisering. Et svært ugjennomsiktig uttrykk vil ikke kunne tolkes «rett» med mindre det fins et allerede oppretta konvensjonelt samband mellom det ugjennomsiktige uttrykket og dets innhold.

- **Semantisk anomaliseringsgrad** er en paraplyterm for betydningsmessige aspekter som bidrar til å gjøre en sammensetning mer eller mindre gjennomsiktig. To viktige og mye diskuterte faktorer for semantisk gjennomsiktighet er *motivasjon* og *komposisjonalitet*. Altså utgjøres den semantiske anomaliseringsgraden delvis av i hvilken grad det finnes en relativt motivert og komposisjonell sammenheng mellom sammensetninga og dens komponenter.

- **Formell anomaliseringsgrad** er en paraplyterm for aspekter som bidrar til at en sammensetning avviker fra regelmessige morfofonologiske mønstre for sammensatte ord i norsk. For eksempel er det en regelmessighet at sammensetninger følger samme bøyningsparadigme som etterleddet, at for- og etterledd er fullstendige i den sammensatte formen, at stammene som inngår i sammensetninga, uttales noenlunde likt innenfor og utenfor den sammensatte konteksten, og at forleddet står ubøyd. Avvik fra dette vil bidra til å øke sammensetningas formelle anomaliseringsgrad.

- I tråd med Svanlund (2002) betegner **motivasjon** her en sammenheng mellom betydninga til sammensetninga og dens komponenters betydning. Dersom det er en tydelig og intuitiv sammenheng mellom et ledds betydning innenfor og utenfor den

sammensatte konteksten, vil leddet kunne sies å være relativt motivert – det er med andre ord intuitivt for en språkbruker hvorfor akkurat dette leddet inngår i sammensetninga. Både forledd, etterledd og sammensetninga som helhet har sin egen motivasjonsstatus. I *fengselsfugl* er for eksempel forleddet vesentlig mer motivert enn etterleddet, mens det motsatte er tilfellet i *blånekte*. I visse tilfeller er begge ledd relativt motiverte, uten at sammensetninga som helhet framstår fullt ut motivert likevel. For eksempel er alle komponenter i *blåhval* motiverte, men det er like fullt ikke slik at alle blåaktige hvaler kan betegnes som *blåhvaler*. Dette fordi det fins noen semantiske tilleggselementer i den konvensjonaliserte betydninga til *blåhval* som gjør at sammensetninga som helhet ikke har full motivasjonsandel. På samme måte som for den overordna gjennomsiktigheten knytter motivasjon seg til forholdet mellom en ordform og én bestemt betydning. Ulike betydninger til en sammensetning kan derfor ha ulik motivasjonsgrad.

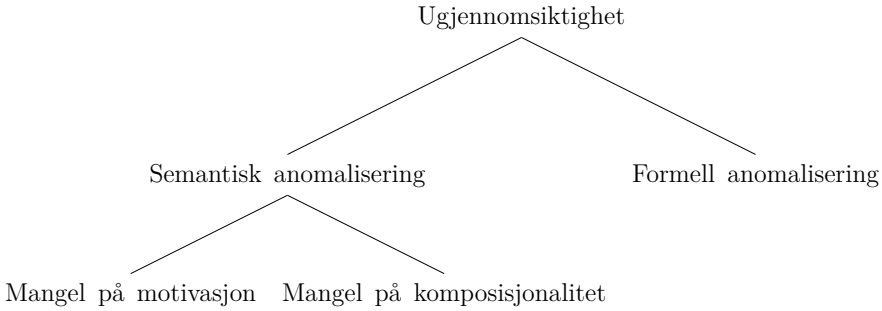
- **Komposisjonalitet** tilskrives ulik vekt og betydning i litteraturen. I denne avhandlinga vil grad av komposisjonalitet være mindre viktig enn for eksempel motivasjon. Dette har sin rot i at jeg – for å skille komposisjonalitet tydelig fra motivasjon og gjennomsiktighet – følger definisjonen til Bundgaard et al. (2006), som åpner for at de aller fleste både gjennomsiktige og ugjennomsiktige sammensetninger er komposisjonelle. Så lenge sammensetninga aktiverer et skjema XY, hvor forleddet spesifiserer et felt innenfor rammen av etterledet, er sammensetninga komposisjonell. Unntaket blir da sammensetninger hvor det er veldig stor begrepslig avstand mellom ledda slik at det blir uklart hva slags felt innenfor Y som spesifiseres av X, som i *makrellfotball*, eller sammensetninger hvor rammen som Y oppretter øyensynlig er irrelevant for en aktuell tolkning, som i *akilleshæl*. Her er det ikke slik at *hæl* i konteksten betyr 'punkt', og at *akilles* betyr 'sårbart', slik at helheten betegner 'et sårbart punkt'. Snarere har sammensetninga mer karakter av å være et minimalt språklig tegn som ikke kan tolkes med utgangspunkt i ledda.

- **Skjematisering** betegner på mange måter motsatsen til anomalisering. Et ord som er maksimalt anomalt, er samtidig minimalt skjematisk. Norske sammensetninger følger i varierende grad regelmessige semantiske og morfologiske mønstre knytta til enten sammensetting generelt, eller en undergruppe en gitt sammensetning inngår i. Som omtalt i delkapittel 2.3 og i underkapittel 2.5.1, fins det regelmessigheter knytta til sammensetninger på ulike språklige nivåer, som gjør at vi kan anta at visse form–betydning-par er svært tilgjengelige for en gjennomsnittlig språkbruker selv om vedkommende ikke er tidligere eksponert for den aktuelle formen.

- **Frekvens** betegner i denne avhandlinga antall forekomster per en viss størrelsesorden, for eksempel per million ord. Såleis indikerer frekvenstall en andel framfor et antall, til tross for at det ikke er helt uvanlig å bruke *frekvens* også om antall

forekomster (se for eksempel Rummelhoff og Frøslie (2023)). Jeg vil likevel avstå fra å benytte *antall* og *frekvens* om hverandre på denne måten. For det første er det uheldig siden frekvens er et mål på hyppighet (noe Rummelhoff og Frøslie (2023) også påpeker). Hvis man teller antall forkomster, sier det ingenting om et fenomens hyppighet, bare hvor mange observasjoner man har. For det andre er *frekvens* som 'antall forekomster per målenhet' rådende innenfor mange fagområder, som f.eks. medisin og fysikk (se Kåss (2020) og Bøe (2020)). For det tredje gjør sammenblandinga av frekvens og antall formen *frekvens* tvetydig, hvilket krever at man må spesifisere om man til enhver tid snakker om *absolutt* eller *relativ* frekvens.

Det er viktig å presisere at summen av morfologisk og semantisk anomalisering i henhold til definisjonene i denne avhandlinga tilsvarer en gjennomsiktighets*prediksjon*. Skillet mellom anomalisering og gjennomsiktighet blir da at sistnevnte betegner den reelle opplevelsen til språkbrukere, mens førstnevnte er en prediksjon av hva denne opplevelsen vil være, basert på målbare egenskaper ved sammensetninga. Termene ovenfor står i en relasjon til hverandre som visualiseres i følgende begrepstaksonomi:



2.7.2 Utsyn

Noe som er helt tydelig, er at leksikografisk metode generelt og leksikografisk seleksjon av sammensetninger spesielt er multivariate prosjekter. For å identifisere de mest etablerte sammensetningene må man på den ene siden måle utbredelse på en tilfredsstillende måte. Utbredelsen kan nemlig fortelle oss hvilke sammensetninger som har høyest diffusjon og repetisjonsfrekvens, som i sin tur indikerer konvensjonalisering og innprenting. Som vi har sett i delkapittel 2.6, bør utbredelse måles ved hjelp av flere korpusvariabler. På den andre siden fins det etter alt å dømme også veletablerte sammensetninger som er konvensjonalisert og innprenta gjennom mer kvalitative mekanismer, som usualisering eller anomalisering. For å gripe tilbake til problemstillinga kan vi foreløpig si at en betingelse for et gunstig sammensetningsutvalg er at det utelukkende inneholder sammensetninger som på et eller annet nivå er etablerte, og at det også fanger sammen-

setninger som er etablerte på ulike vis. Dette fordrer at man benytter seleksjonsmetoder som bygger på multivariate forståelser av både de kvantitative og de kvalitative sidene av etablering.

I neste kapittel vil jeg konkretisere framgangsmåten i delstudiene i avhandlinga, før jeg oppsummerer bakgrunn og funn for hver delstudie og seinere diskuterer problemstillinga i lys av disse og de teoretiske avveiningene som er presentert i inneværende kapittel.

Kapittel 3

Framgangsmåte

I dette kapittelet redegjør jeg for den metodiske framgangsmåten i de tre delstudiene og i prosjektet som helhet.

3.1 Metoder

3.1.1 Semantisk analyse

I delstudie 1 benyttes en kvalitativ innfallsvinkel for å vurdere den semantiske anomaliseringsgraden til et utvalg sammensetninger. Til dette formålet utføres en semantisk analyse basert på fem anomaliseringsfaktorer fra sammensetningslitteraturen (se detaljert omtale i delkapittel 4.1) som belyser ulike semantiske egenskaper ved sammensatte ord. Disse faktorene blir i delstudien operasjonalisert slik at en sammensetnings skår på hver faktor sammenlagt gir en indikasjon på sammensetningas totale semantiske anomalisering. For tre av faktorene brukes Bokmålsordboka og subsidiært Det Norske Akademis ordbok (NAOB) som fasit på hva den mest konvensjonelle betydninga til en sammensetning og dens ledd er. Samtlige av disse faktorene dreier seg om forholdet mellom sammensetningsleddas betydning og helhetens betydning. For de to resterende faktorene benyttes LBK for å undersøke forhold mellom en aktuell sammensetning og andre sammensetninger med samme for- eller etterledd. Begge disse faktorene dreier seg om hvorvidt en aktuell sammensetning følger et etablert semantisk mønster.

3.1.2 Korpus

Korpus som forskningsressurs ble kort introdusert i delkapittel 1.4. Ifølge Stefanowitsch (2020, 21) har korpusbaserte språkstudier minst 100 års fartstid, men ikke uten visse opphold underveis (der bruksbaserte tilnærminger hadde relativt lav status). Det er

bare relativt nylig at korpusbaserte tilnærminger har inntatt en sentral posisjon innenfor lingvistisk metode. Dessuten påpeker Stefanowitsch at en vanlig kritikk mot korpuslingvistikk er at korpuset som objekt nødvendigvis gir en mangelfull framstilling av sitt målobjekt (se dessuten diskusjon i underkapittel 2.6.1). Stefanowitsch kontrer denne kritikken med at så godt som alle forskningsdata, innenfor enhver vitenskap, er mangelfulle. Forskeren og forskningas oppgave er å utlede teorier og teste hypoteser på bakgrunn av de mangelfulle dataene.

De fleste korpusstudier, så vel som delstudie 2 og 3, er undersøkelser av tendenser og probabiliteter (Mehl, 2016; Stefanowitsch, 2020, 68). En av tendensene som undersøkes ved hjelp av korpus i denne avhandlinga, er overordna sett hvilke sammensetninger som viser høy etableringsgrad gjennom høy ususutbredelse i norsk, og som dermed framstår som gode ordbokkandidater ut fra et reint kvantitativt perspektiv.

3.1.3 Korpusbaserte kvantitative metoder

I denne avhandlinga benyttes flere korpusbaserte metoder. I artikkel 1 benyttes korpusfrekvens i forbindelse med at jeg bedømmer anomaliseringsgraden til et knippe sammensetninger, herunder hvilke skjematiske mønstre som fins blant sammensetninger med samme for- eller etterledd som sammensetninga under vurdering (se detaljert omtale i delkapittel 4.1). Delstudie 2 søker å finne svar på hvordan korpus kan utnyttes slik at det gir mest mulig valide estimater av sammensetningers etableringsgrad. Til dette formålet evalueres fem ulike korpusmål via kryssvalidering (se omtale i neste avsnitt). Delstudie 3 evaluerer i hvilken grad korpusmålene frekvens og spredning predikerer søkeinteresse. Her er frekvens konfigurert som antall forekomster delt på korpusets størrelse (se omtale i underkapittel 2.6.3, mens spredning er konfigurert som snittet av spredningsmålene *Deviation of Proportions* (Gries, 2008) og *Juillands D* (Juilland et al., 1970) beregna over metavariablene domene og år i LBK. Med denne konfigurasjonen fordeler spredningsverdiene seg på en skala fra 0 til 1, der høye verdier indikerer jevn spredning. Spredningsmålet indikerer altså hvor jevnt utspredd en sammensetnings forekomster er med hensyn til sjanger og tid. Visse sjangerspesifikke ord kan dermed få relativt jevn spredning om deres forekomster fordeler seg jevnt med hensyn til tid, og vice versa. Spredning konfigureres videre med utgangspunkt i *Deviation of Proportions* aleine i samband med at jeg foreslår en konkret prosedyre for utvelgelse av sammensetninger i kapittel 5.

3.1.4 Kryssvalidering

Kryssvalidering er en form for intern validering av estimater (Pripp, 2020). Gitt en statistisk modell anvendt på et datasett kan man bruke kryssvalideringsteknikker for

ytterligere testing av modellen, uten innsamling av nye data. Dette kan være svært nyttig i kontekst av statistiske beregninger gjort på et korpus. Om man for eksempel har beregna frekvensen til leksemet *besudle* i LBK til 1,22 forekomster per million ord, så veit man ikke hvor godt dette estimatet er, altså i hvilken grad det stemmer at *besudle* faktisk forekommer 1,22 ganger per million ord i usus. Validiteten til den statistiske modellen FREKVENS anvendt på et korpusmateriale avgjøres av hvor treffsikkert modellen gjenspeiler frekvens i usus. Vi har som nevnt i kapittel 2 ikke tilgang på hele usus, derfor må vi nøye oss med å studere validiteten ved hjelp av data vi har tilgang på. Valget står da mellom å finne andre data som kan støtte eller svekke estimatet i modellen, eller å skille ut deler av korpuset for så å anvende det som om det var eksterne data.

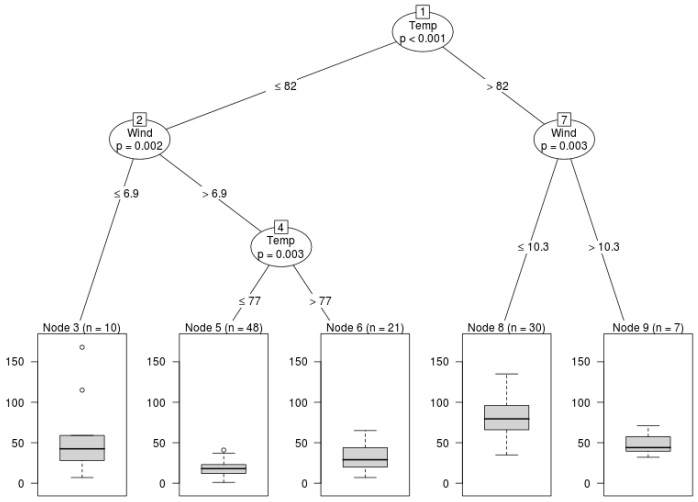
Det sistnevnte er et viktig steg i en kryssvalideringsprosedyre. Her kan for eksempel et korpus splittes i to deler, hvorav den ene utgjør et såkalt treningssett mens den andre utgjør testsettet. Disse to datasetta simulerer henholdsvis et utvalg og en populasjon. Målinger, eller estimater, gjort på henholdsvis trenings- og testsettet blir så sammenlikna, hvorpå differansen i eksempelvis frekvens indikerer hvor presis (og valid) frekvensestimatet i treningssettet er. Prosedyren kan repeteres mange ganger med forskjellige uttrekk fra korpuset som testsett, slik at resultatet ikke blir influert av idiosynkratiske trekk ved de ulike uttrekka.

Om det er stort avvik mellom estimata som gjøres på bakgrunn av treningssetta og de korresponderende testsetta, vitner dette om at datasettet, for eksempel korpuset, har høy variabilitet i verdiene av den variabelen som måles, for eksempel frekvensen til *besudle*. At det varierer mye hvor frekvent *besudle* forekommer i ulike korpusdeler, kan selvfølgelig være en følge av at *besudle* har en svært ujevn distribusjon i usus. Men det er også fullt mulig at det er en arbitrær effekt av ett eneste eksemplar i korpuset, at det for eksempel inngår en tekst av en forfatter som har lagt sin elsk på dette ordet og derfor bruker det abnormalt ofte. Uansett indikerer høy variabilitet at det er stor usikkerhet om reliabiliteten, altså at man får veldig forskjellige tall avhengig av hva slags materiale man ser på. Dette fører videre til at den statistiske modellen har lav validitet; om for eksempel frekvensen fra én del av korpuset er systematisk uegna til å gjenspeile frekvensen i en annen del, er det svært usikkert om korpusfrekvensen gir et presist estimat av ususfrekvensen (se for øvrig diskusjonen om korpusfrekvens og variabilitet i 2.6.3). Kryssvalidering anvendes i artikkel 2 for å måle validiteten og reliabiliteten til frekvens og fire ulike spredningsmål som mål på utbredelse i usus.

3.1.5 Inferenstrær og randomiserte skoger

Om man ønsker å predikere fordelinga i en variabel basert på for eksempel frekvens- eller spredningstall fra korpus, er det viktig å ta i betraktning at slikt tallmateriale som

regel ikke er normalfordelt (se detaljert omtale i 2.6.6). Ikke-normalfordelte data krever ikke-parametriske analyser. I en korpussammenheng vil dette for eksempel si at man gjør analyser som tar utgangspunkt i leksemenes frekvens*rangeringer* framfor deres faktiske frekvensverdier i utvalget. To ikke-parametriske, sammenhørende analysemetoder som i økende grad blir anvendt på lingvistiske korpusdata, er *inferenstrær* og *randomiserte skoger* (se f.eks. Levshina (2015, 2020); Tagliamonte og Baayen (2012); Winter (2020)).



Figur 3.1: Eksempel på inferenstre med prediktorene temperatur (Temp) og vindstyrke (Wind) og responsvariabelen ozonnivå. Y-aksen i boksdiagramma viser ozonnivå.

Utputten til en inferenstreanalyse er, som illustrert i figur 3.1,[1] en treliknende grafisk framstilling med ett eller flere forgreiningspunkter (noder). I dette eksempelet fins det to prediktorer (eller uavhengige variabler), temperatur (Temp) og vindstyrke (Wind), og responsvariabelen (eller den avhengige variabelen) ozonnivå. Målet er altså å finne ut hvordan konsentrasjonen av ozon i lufta i et gitt miljø påvirkes av temperatur og vindstyrke. Nederst i treet ser vi fem boksdiagram. Disse representerer fem distinkte underkategorier av datasettet, og visualiserer fordelinga av ozonnivå innenfor hver underkategori. Det er altså responsvariabelen som utgjør y-aksen i boksdiagramma, og i det aktuelle eksempelet ser vi hvordan underkategorier av observasjoner av ozonnivå fordeler seg basert på kombinasjoner av temperatur- og vindstyrkeverdier. Over boksdiagramma angis antall observasjoner i hver underkategori (f.eks. $n = 10$).

For å lage et inferenstre må man spesifisere en grenseverdi for assosiasjonsstyrken mellom prediktorene og responsvariabelen, og et minstemål på antall observasjoner for hver

[1] Eksempeltreet er henta fra <datawookie.dev/blog/2013/05/package-party-conditional-inference-trees/>.

underkategori i treet. Treet blir til ved at en rekursiv algoritme måler hvilken prediktor som har den sterkeste assosiasjonen med responsvariabelen. Først lages det to versjoner av responsvariabelen, den originale og en tilfeldig permutert versjon. Den permuterte versjonen tjener som en nullhypotese i og med den har en arbitrær relasjon mellom prediktorverdiene og de permuterte responsverdiene. Deretter blir det beregna en statistisk observator med tilhørende p-verdi for styrken i assosiasjonen mellom responsvariabelen og hver prediktor. Prediktoren som har lavest p-verdi, og dermed den sterkeste assosiasjonen med responsvariabelen, danner så grunnlag for den øverste forgreininga i treet (se Levshina (2020) for flere detaljer om beregninga av assosiasjonsstyrke). Prosedyren gjentas så på hver av de to forgreiningene inntil spesifiserte stoppbetingelser er oppfylt. For kontinuerlige prediktorer baseres forgreininger på grenseverdien i prediktoren som maksimerer forskjellen i responsverdier for de to resulterende underkategoriene. Grenseverdien identifiseres gjennom at algoritmen tester alle mulige dikotome oppdelinger av prediktoren.

I treet i figur 3.1 ser vi øverst i Node 1 at datasettet splittes i to underkategorier basert på prediktoren temperatur. På høyre side av treet fins observasjoner med temperatur høyere enn 82 (fahrenheit), og på venstre side fins de resterende observasjonene, altså temperatur lik eller lavere enn 82. Algoritmen foretar så ytterligere oppdelinger basert på hvilken prediktor som er sterkest assosiert med responsvariabelen innenfor temperaturunderkategoriene. I eksempeltreet ser vi i Node 2 og 7 at hver av underkategoriene splittes i henhold til vindstyrke, før en av disse resulterende underkategoriene blir ytterligere splitta basert på temperatur.

Treet i figur 3.1 gir oss en intuitiv grafisk framstilling av forholdet mellom henholdsvis vindstyrke og temperatur på den ene siden og ozonnivå på den andre. Som man kan se, har boksdiagrammet i Node 8 den høyeste medianen hva gjelder ozonnivå. Om vi sporer forgreiningene ovenfor denne noden, kan vi se at underkategorien den representerer, inneholder observasjoner med temperatur over 82 grader fahrenheit og vindstyrke under 10,3 m/s. Treet indikerer med andre ord at en tilstand med høy temperatur og moderat vindstyrke er forbundet med jevnt over høyere ozonnivåer enn for eksempel en tilstand med høy temperatur og høy vindstyrke. Det er en styrke ved inferenstrær at de på et intuitivt sett visualiserer hvordan en kombinasjon av prediktorer, som for eksempel temperatur og vindstyrke, predikerer variasjon i responsvariabelen.

Randomiserte skoger blir av blant andre Levshina (2015, 2020) og Tagliamonte og Baayen (2012) beskrevet som en skog av inferenstrær. Utputten til skogmetoden er, til forskjell fra inferenstremetoden, et diagram som viser den relative betydninga (importance) til hver av prediktorene. For å lage en randomisert skog spesifiserer man først hvor mange inferenstrær skogen skal bestå av. Om skogen består av 1000 trær, kjøres prosedyren beskrevet ovenfor 1000 ganger basert på 1000 tilfeldige uttrekk fra datasettet,

med tilfeldige konstellasjoner av prediktorer. Dette innebærer at hver enkelt prediktor er utelatt i enkelte trær. Deretter beregnes gjennomsnittlig avstand fra den permuterte responsvariabelen, altså nullhypotesen, for hver prediktor over samtlige trær. Denne avstanden benyttes til slutt til å kalkulere den relative betydninga til hver prediktor.

Tre særlig viktige grunner til at inferenstrær og randomiserte skoger er gunstige å anvende og kombinere i inferensielle studier av korpusdata, er at de takler nær korrelasjon mellom prediktorer, et stort antall prediktorer relativt til antall observasjoner og ikke-linearitet (Levshina, 2020). I et design som i artikkel 3, der flere av prediktorene er korpusestimater av bruksmønstre, herunder frekvens og spredningstall fra samme korpus, vil man ofte ha nær korrelasjon mellom prediktorer fordi de vanligste orda i språket (eller datautvalget) vil ha systematisk høyere frekvens- og spredningsskår enn de sjeldne, nesten uavhengig av hvordan man måler utbredelse. Om et ord har veldig mange forekomster i korpuset, er det sannsynlig at det er frekvent på tvers av ulike korpusdeler, og at det derfor også har jevn spredning. Videre er det en styrke at man kan inkludere et stort antall numeriske og/eller kategoriske variabler i slike studier, også i de tilfeller man har relativt få observasjoner. Om man for eksempel er interessert i hva som påvirker vekslinga mellom to likeverdige, men forholdsvis sjeldne syntaktiske konstruksjoner, kan det være relevant å kontrollere for mange lingvistiske og ekstralingvistiske variabler. Da vil man ha få observasjoner, men mange prediktorer. Til sist er det uproblematisk å bruke de nevnte metodene i kontekster der det er et ikke-linært forhold mellom prediktorer og responsvariabelen, hvilket vil si at verdien til observasjonene i en eller flere prediktorer ikke øker proporsjonalt med observasjonene i responsvariabelen. Delstudie 2 inneholder for eksempel en framstilling som viser at det blant annet dannes en s-liknende kurve mellom korpusfrekvens og antall korpusdeler et ord forekommer i.

3.2 Materiale

Forskningsspørsmåla i denne avhandlinga knytter seg til (den stadig voksende) populasjonen av sammensatte ord i norsk, og hvordan man skal håndtere denne populasjonen i leksikografiske prosjekt. Noe som er gitt a priori, er at leksikografer må gjøre et utvalg, altså at det kun er et utvalg av denne populasjonen som skal innlemmes i ordbøker. Akkurat som det er upraktisk og prinsipielt uoppnåelig å innlemme hele sammensetningspopulasjonen i ordbøker, er det upraktisk og prinsipielt uoppnåelig å anvende hele sammensetningspopulasjonen som datagrunnlag i undersøkelser av den. Også leksikologiske undersøkelser må baseres på datautvalg. I det følgende skildres utvalga som er brukt i denne avhandlinga.

3.2.1 Leksikografisk bokmålskorpus

Utover de ulike sammensetningsutvalga (se underkapittel 3.2.2 tjener Leksikografisk bokmålskorpus (LBK) som datagrunnlag for å gjenspeile bruken av sammensetningene. Korpuset består av rundt 100 millioner ord og inneholder språkmateriale fra perioden 1985–2013 (Fjeld et al., 2020). Korpusmaterialet er såkalt «balansert», hvilket vil si at den sjangermessige fordelinga er ment å være proporsjonal med hva gjennomsnittlige språkbrukere blir eksponert for av forskjellige skriftlige sjangre. I LBKs tilfelle ligger statistiske data om folks lesevaner til grunn for hvor stor andel av korpuset som utgjøres av domena avistekster, sakprosa, skjønnlitteratur, tv-tekst og upublisert materiale. Likevel er det ikke et en-til-en-forhold mellom fordelinga av ulike tekstdomener i folks lesevaner og i korpuset. For det første var den aktuelle undersøkelsen av lesevaner svært underspesifisert. For eksempel er 35 % av folks lesing knytta til «internett», hvilket gir liten informasjon om hva slags tekster det er snakk om. For det andre var LBK begrensa av at ikke alle teksttyper av rettighetsgrunner var like lette å samle. For det tredje støtter også LBK-redaksjonen seg på fordelinga i andre store «balanserte» korpus for å komme fram til den endelige fordelinga i LBK, som domenemessig består av 20 % periodika, 45 % sakprosa, 25 % skjønnlitteratur, 5 % tv-tekst og 5 % upublisert materiale (Fjeld et al., 2020, 107). Disse tekstkategoriene er videre delt inn i mer spesifikke underkategorier som regionalaviser, lærebøker, lyrikk, teksting av tv-serier og blogger.

Man kan si at LBK representerer en viss bredde av det gjennomsnittlige språkbrukere blir eksponert for av skriftlige tekstkategorier på norsk bokmål. Den fulle bredden er imidlertid ikke representert, da for eksempel en sentral sjanger som personlig korrespondanse per e-post, tekstmeldinger eller brev er fraværende. Dessuten forsøker ikke LBK å fange gjennomsnittlig tekstproduksjon. Representativiteten som fins, gjelder først og fremst typenivået. På eksemplarnivå fins det noen avveininger knytta til tekstforfatternes kjønn, alder og sosiale tilhørighet, men det er vanskelig å vite om dette har noen reell betydning for hvor typiske eller representative teksteksemplara er for sine typer.

LBK anvendes i samtlige av avhandlingas tre delstudier pluss i kapittel 5. I studie 1 utgjør det et datagrunnlag for to av variablene i en anomaliseringsberegning, mens det i studie 2 og 3 utgjør datagrunnlaget for beregning av frekvens- og spredningsestimater. I kapittel 5 benyttes det i et forslag til prosedyre for seleksjon av sammensetninger. Mer detaljerte beskrivelser fins i selve artiklene, i sammendraga i kapittel 4, og i beskrivelsen av prosedyren i kapittel 5.

3.2.2 Sammensetningsutvalg

I delstudiene i denne avhandlinga brukes det bestemte sammensetningsutvalg. Utvalga varierer basert på hva slags undersøkelse som foretas i hver delstudie, men alle utvalga har en viss grad av ordklassevariasjon i sammensetningsledda. Hvert enkelt utvalg beskrives nærmere i det følgende.

Utvalget i delstudie 1 består av 79 sammensetninger med førsteleddene *svart-*, *tanke-* og *vandre-*. Dette datasettet er valgt fordi det fanger mange typiske problemstillinger knytta til bedømmelsen av den semantiske anomaliseringsgraden til ulike sammensetninger. Det adjektiviske forleddet *svart-* er kjennetegna av polysemi. NAOB lister åtte forskjellige betydninger av adjektivet *svart* (per 27.11.2023), hvorav flere betydninger deles opp i videre underbetydninger og faste uttrykk. De ulike definisjonene indikerer at *svart* har mange aktuelle metaforiske utbygginger, og disse gjør seg gjeldende også i sammensetninger med *svart-* som forledd.

Videre er sammensetninger med forleddet *tanke-* kjennetegna av et abstrakt eller uhåndgripelig betydningsinnhold. Intuitivt virker det overkommelig å beskrive relasjonen mellom sammensetningsledd som betegner fysiske objekter; en jerngryte er for eksempel en gryte som består av jern. Med abstrakte ledd blir det mer problematisk å avgjøre hva relasjonen består i, da den ofte må baseres på billedlige forståelser av fenomener. Med en sammensetning som *tankerekke* går det selvfølgelig an å tenke seg en rekke som består av tanker, men dette kan kun stemme på et billedlig nivå der tanker kan deles opp og kvantifiseres og deretter arrangeres på rekke. Forholdet mellom ledda i *tankerekke* er derfor ikke like selvsagt som i *jerngryte*.

Til sist er sammensetninger med forleddet *vandre-* kjennetegna av at det fins forskjelligarta relasjoner mellom for- og etterleddet. Siden forleddet er et intransitivt verb, er det noenlunde forutsigbart at en viss andel etterledd spesifiserer agens til verbhandlinga. Dette er for eksempel tilfellet i sammensetningene *vandremaur*, *vandreutstilling* og *vandrenyre*. I mange tilfeller synes forleddet *vandre-* imidlertid å ha en avbleika betydning. Det er for eksempel vanskelig å tenke seg at utstillinger og nyrer skal vandre i betydning 'ferdes til fots' (se NAOB). Dessuten er det flere tilfeller av etterledd som ikke spesifiserer agens til verbhandlinga, sånn som for eksempel i *vandretur*. Altså fins det en viss grad av relasjonspolysemi blant *vandre*-sammensetningene.

Sammensetningsutvalget i delstudie 1 inneholder dermed polyseme, metaforiske og abstrakte ledd, og relasjonspolysemi. Dette er ikke uvanlige egenskaper i sammensetningspopulasjonen, og metafor og polysemi gir ofte opphav til semantisk anomali i enkeltsammensetninger, noe også analysen i delstudie 1 viser (se 4.1).

I delstudie 2 benyttes et utvalg med 273 forskjelligarta sammensetninger. I denne del-

studien gjøres det kun kvantitative analyser, altså er det kun distribusjonsvariasjonen i utvalget som er relevant. Som nevnt i underkapittel 3.1.5 har leksemfrekvenser gjerne en zipfisk distribusjon der frekvensrangeringa til leksemene er omvendt proporsjonal med frekvensestimatet. Selv om det er vanskelig å si noe definitivt om størrelsesforholdet mellom et mellomstort korpus som LBK, som inneholder ca. 100 millioner ord, og sammensetningspopulasjonen i usus, er det svært sannsynlig at sistnevnte har en zipfisk distribusjon på lik linje med et mellomstort korpus (se f.eks. Piantadosi (2014)).

Sammensetningsutvalget i delstudie 1 inngår også i utvalget i delstudie 2. I dette utvalget er det mange lavfrekvente sammensetninger, derfor har materialet i artikkel 2 blitt supplert med en del mer frekvente sammensetninger med forledda *år-*, *fram-*, *bak-*, *med-*, *arbeid-*, *om-*, *under-* og *over-*. Disse sammensetningene er håndplukket basert på frekvensestimater fra LBK. Totalt inneholder utvalget dermed et mangfold av ulike distribusjoner, og spenner fra svært høyfrekvente sammensetninger til hapax legomena. Selv om dette sannsynligvis ikke gir en balansert representasjon av sammensetningspopulasjonen, gir det grunnlag for å si noe om styrker og svakheter i prediksjonsevnen til ulike korpusmål i møte med ulike distribusjoner, hvilket er det overordna formålet med studien. Dessuten har korpusfrekvensen til sammensetningsutvalget en zipfisk fordeling på lik linje med leksemfrekvenser i usus.

I delstudie 3 består utvalget av sammensetninger som enten er ordbokførte, korpusbelagte eller oppslåtte (se underkapittel om søkestatistikk nedenfor) innenfor fem alfabetiske strekk som sammenfaller med redigeringsbolker i revisjonen av *Bokmåls-* og *Nynorskordboka*.[2] Utvelgelsen beskrives i detalj i artikkel 3. Hensikten med å bruke redigeringsbolker som datautvalg er at studien dermed tar utgangspunkt i en materie som leksikografer ofte hanskes med, nemlig en grafemisk bestemt samling av potensielle ordbokkandidater. De konkrete redigeringsbolkene som er valgt, er dessuten redigert forholdsvis nylig på bakgrunn av blant annet utbredelse i LBK.

3.2.3 Søkestatistikk

En av fordelene ved elektroniske ordbøker er at man kan monitorere søkevirksomheten til brukerne. I BOB og NOB genererer hvert søk en url, dette gjelder også om brukeren klikker seg inn via en ekstern søkemotor. Url-ene loggføres slik at man ved hjelp av programmeringsverktøy kan trekke ut tallmateriale knytta til blant annet hvor mange ganger ulike søkeuttrykk har blitt benytta innenfor ulike tidsperioder. Slike søkestatistikker er en verdifull kilde til brukernes søkeatferd, hva de interesserer seg for, og hvilke behov de har. For eksempel kan søkestatistikken indikere at brukere i utstrakt grad søker etter verbalsubstantiv, flerordsuttrykk og sammensetninger, noe som kan gi ordbokre-

[2]Se informasjon om Revisjonsprosjektet på <https://www.uib.no/lle/revisjonsprosjektet>.

daksjonen grunn til å prioritere slike konstruksjoner.

I delstudie 3 benyttes variabelen *oppslagsregularitet*, som er konfigurert som antall loggførte søk på et leksem multiplisert med spredninga til disse søkene på tvers av de fem årene 2016–2020, som søkestatistikken i delstudie 3 er henta fra. Spredninga er i denne sammenhengen operasjonalisert gjennom snittet av DP og Juillands D (se forklaring av disse i delkapittel 3.1.3). Oppslagsregulariteten indikerer med andre ord hvor mange loggførte søk et sammensetningsleksem har, kontrollert for hvor jevnt disse søkene fordeler seg over tid.

Det er imidlertid viktig å ha med seg at ikke all aktivitet knytta til BOB og NOB loggføres. For eksempel kan en bruker som er ute etter et bestemt ord, søke i en ekstern søkemotor og ut fra manglende treff fra ordbøkene der, slutte at ordet ikke står i ordbøkene. Likeleis kan en bruker begynne å skrive sitt søkeuttrykk i ordbøkenes grensesnitt og følge med på hva slags ord som foreslås i rullegardinmenyen som dukker opp under søkefeltet. Om ordet søkeren er ute etter, ikke dukker opp i denne menyen, eller om brukeren bare er interessert i stavemåten og dermed får svar fra det som kommer i menyen, kan brukeren avslutte aktiviteten sin uten at det blir registrert som et søk.

Kapittel 4

Sammendrag av delstudier

4.1 Delstudie 1

Delstudie 1 utgjøres av en vitenskapelig artikkel ved navn «Svartsjuk tankelesing på vandresafari – en modell for bedømmelse av sammensatte ord gjennomsiktighet» som ble publisert i tidsskriftet *LexicoNordica* i 2020.

4.1.1 Kontekst og bakgrunn

Gjennomsiktighet har tradisjonelt vært anvendt som et viktig kriterium for å skille mellom gode og dårlige ordbokkandidater innenfor den store mengden av sammensatte ord i norsk (Fjeld & Vikør, 2008, 160). Likevel har jeg ikke klart å finne studier som fullt ut spesifiserer hvilke faktorer som er relevante når man forsøker å skjelne mellom gjennomsiktige og ugjennomsiktige sammensetninger, og hvordan man kan gå fram for å gjøre disse vurderingene i en leksikografisk setting. I delkapittel 2.5 gjennomgås et utvalg av de teoretiske og empiriske studiene som vedrører den betydningsmessige sammenhengen mellom sammensetninger og deres komponenter. Selv om det er stor variasjon i termbruk og innfallsvinkel, blir det til sammen et anselig fundament å basere en semantisk anomaliseringsmodell på.

Det er først og fremst den leksikografiske settingen som motiverer formuleringa av en ny modell. Siden det er gjort mengder med forskning på gjennomsiktigheten til sammensatte ord (se blant annet Schäfer (2018) for en oversikt), og siden for eksempel Downing (1977) og Borque (2014) spesifiserer hvilke konkrete faktorer som påvirker denne gjennomsiktigheten, er det diskutabelt om det å formulere enda en modell gir noen nye innsikter til spørsmålet om gjennomsiktighet i sammensetninger. Jeg vil imidlertid argumentere for at det går en viktig skillelinje mellom hva som bidrar til å bekrefte eller avkrefte hypoteser avfødt av lingvistisk teori på den ene siden, og hva som bidrar til å løse prinsipielle

og praktiske problemstillinger innenfor leksikografi på den andre. Der en ny anomaliseringsmodell muligens gir et begrensa bidrag til det førstnevnte, kan den gi et vesentlig bidrag til det sistnevnte. Ambisjonen om å løse bestemte praktisk–prinsipielle problemstillinger spiller dessuten en stor rolle for utforminga av den aktuelle modellen. Dersom sammensetninger skal vurderes enkeltvis, er det et særlig krav til en leksikografisk modell at den ikke er for tidkrevende å bruke, hvilket i sin tur begrenser antallet variabler det er hensiktsmessig å inkludere. I det følgende forklares kort bakgrunnen for de aktuelle variablene i studien.

I modellen i delstudie 1 inkluderes totalt fem variabler, hvorav tre bygger på prinsipielle argumenter og empiriske funn hos Svanlund (2002, 2009), én bygger på empiriske funn hos Loenheim (2019) og til dels Svanlund (2009), og én bygger på en serie psykolingvistiske eksperimenter av Gagné og Shoben (1997); Maguire, Devereux, Costello og Cater (2007); Maguire, Maguire og Cater (2010); Shoben og Gagné (1997). Variablene og deres operasjonaliseringer i modellen vektlegger henholdsvis følgende aspekter ved sammensetningers semantiske anomalisering: 1) motivasjonsandel (i artikkel 1 kalt *tilleggselementer*), 2) motivasjonsgrad, 3) intern disambiguering, 4) skjematiske forbilder (i artikkel 1 kalt *analogi*) og 5) skjematisk produktivitet (i artikkel 1 kalt *relasjonsmønster*).

Svanlund (2002) påpeker at opplevelse av gjennomsiktighet bare kan måles ved den første eksponeringa for et uttrykk. Etter dette vil det være usikkert om språkbrukerens tolkninger er basert på skjematiske strukturer og språklig minne knytta til ledda, eller om de først og fremst er basert på språklig minne knytta til hele sammensetninga som en enhet. I etterhånd har vi derfor ikke tilgang på opplevd gjennomsiktighet, men må ta til takke med å gjøre retrospektive bedømmelser. Modellen i delstudie 1 er derfor utforma med tanke på bedømmelser av semantisk anomaliseringsgrad, hvilket etter alle solemerker har stor påvirkning på opplevelse av gjennomsiktighet. Siden flere av bedømmelseskriteriene har et empirisk fundament, er det likevel trolig at bedømmelsen til en viss grad svarer på spørsmålet «hvor sannsynlig er det at en tidligere ueksponert språkbruker av sammensetning *a* tolker den i betydning *b*».

I delstudie 1 brukes motivasjon som en faktor for anomaliseringsgrad. Et ubesvart spørsmål knytta til denne variabelen er hvilken rolle polysemi spiller for motivasjon. I studien analyseres for eksempel mange sammensetninger med forleddet *svart-*, hvilket er usedvanlig polysemt. Som en forenkling anses *svart* i studien å ha én tilgjengelig standardbetydning for henholdsvis substantivet og adjektivet, nemlig ‘svart farge’ og ‘som har svart farge’ (se BOB). Dersom en sammensetning bygger på andre betydninger av *svart*, gjør dette sammensetninga mindre motivert. Denne løsninga åpner ikke for at et ord kan ha flere relativt tilgjengelige betydninger, og at ulike betydninger kan ha ulik grad av mental tilgjengelighet i ulike kontekster. For eksempel er trolig den politiske betydnin-

ga av *blå* svært nærliggende i politiske kontekster, men mindre tilgjengelig i kontekst av et formgivningskurs. Slike omstendigheter blir abstrahert bort fra delstudie 1, men kan med fordel undersøkes nærmere i framtidige studier av begrepet motivasjon.

I delstudie 1 svarer semantisk ugjennomsiktighet til det som ellers i denne avhandlinga omtales som semantisk anomalisering (se begrepsavklaring i underkapittel 2.7.1). Anomaliseringsgrad har trolig en viss korrelasjon med etableringsgrad. I hvert fall kan man tenke seg at sammensetninger som har oppstått og etablert seg i et tidligere språksteg, eller i en annen tidsånd, kan framstå anomale i synkron sammenheng. Det er imidlertid grunn til å presisere at også relativt etablerte sammensetninger kan ha lav anomaliseringsgrad, og at helt nylagde sammensetninger kan ha noenlunde høy anomaliseringsgrad allerede tidlig i etableringsfasen (se blant annet Svanlund (2002) for dette poenget). I tillegg kommer poenget fra Svanlund om at sammensetninger er prinsipielt mangetydige, hvilket innebærer at en sammensatt form kan bli betydelig anomalisert dersom en avvikende betydning knytta til den sammensatte formen usualiseres (se delkapittel 2.2.1). I det følgende oppsummeres hovedfunna i delstudie 1.

4.1.2 Funn og diskusjon

Anomaliseringsmodellen i delstudie 1 rangerer sammensetninger langs en skala. Utvalget som modellen testes på, fordeler seg på sju av skalaens åtte trinn. Ingen sammensetninger vurderes som minimalt anomaliserte i henhold til modellens variabler. Hovedgrunnen til dette er at faktoren *disambiguering* (jf. punkt 3 ovenfor) i veldig få tilfeller gir uttelling.

Det kan være en styrke ved modellen at den gir en relativt finkorna inndeling av sammensetningene med hensyn til anomalisering. Siden anomalisering med all sannsynlighet er et graduelt fenomen, vil en skala med mange trinn gi en bedre representasjon av fenomenet enn for eksempel en binær kategorisering. Mange av sammensetningene som vurderes i delstudie 1 vil trolig kvalifisere for oppslag via etableringsgrad i og med at de har høy utbredelse. I seleksjon av sammensetninger må derfor anomaliseringsgraden ses i sammenheng med andre variabler. Man kan likevel argumentere for at i alle fall de sammensetningene som befinner seg i det mest anomaliserte sjiktet, sannsynligvis vil påkalle en del spørsmål og nysgjerrighet hos brukerne. Dette er trolig en tilstrekkelig grunn til å innlemme disse, uavhengig av andre variabler.

Man kan ikke si noe om validiteten til modellen uten å evaluere dens evne til å predikere resultater fra et psykolingvistisk eksperiment der språkbrukere blir bedt om å rangere tolkbarheten til sammensatte uttrykk de er tidligere ueksponert for (en viss evaluering gjøres likevel i delstudie 3, se delkapittel 4.3 nedenfor). Selv om det er prinsipielt mulig å gjennomføre en slik studie, vil det være svært krevende å finne sammensetninger av varierende etableringsgrad som et noenlunde representativt utvalg språkbrukere

er tidligere ueksponert for. Samtidig vil det være meningsløst å inkludere bare relativt nylagde sammensetninger i en slik studie, da disse typisk ikke har noen etablerte betydninger som kan utgjøre en bunnlinje av «rette» tolkninger. Spørsmålet blir da om stor fortolkningsvariasjon i seg selv er et tegn på gjennomsiktighet.

For å svare på om fortolkningsvariasjon i seg selv indikerer gjennomsiktighet, må man ha i mente at gjennomsiktighet slik det er definert her, dreier seg om forholdet mellom en form og dens betydning. Om et utvalg språkbrukere i møte med samme form finner forskjellige form–betydning-par, viser dette bare at formen kan anta mange betydninger (noe som er symptomatisk for sammensatte former (Svanlund, 2002), se også underkapittel 2.4). Dersom ingen form–betydning-par utpeker seg som dominerende blant språkbrukernes fortolkninger, kan det være tegn på at ingen av para har høy etableringsgrad. Verken variasjon i hva slags betydning språkbrukere attribuerer til en isolert sammensetning, eller fravær av en dominerende fortolkning gir noen direkte indikasjon på gjennomsiktighetsgraden til et eller flere form–betydning-par. Men dersom et gitt form–betydning-par er tilnærmet fraværende i et materiale der en stor gruppe språkbrukere har gitt uttømmende fortolkninger til den aktuelle formen, har man i alle fall grunnlag for å si at det aktuelle paret er forholdsvis utilgjengelig for den aktuelle gruppa språkbrukere. Om 2000 språkbrukere overser betydninga 'klumsete person' til formen *klossmajor*, gir det et visst grunnlag for å si at dette form–betydning-paret er ugjennomsiktig.

En anomaliseringsmodell må nødvendigvis gjøre visse tilpasninger som ikke er fullstendig i tråd med lingvistisk kunnskap for øvrig. I modellen i delstudie 1 fins det helt klart problematiske sider ved faktoren *motivasjonsgrad*. For å regne ut motivasjonsgraden er det vesentlig at man klarer å identifisere den mest etablerte betydninga til ledda når de opptrer utenfor sammensetninger. I delstudien er øverste betydning i BOB (eller ev. NAOB) benytta som indikasjon på hva den mest etablerte betydninga er. Dette er ingen ufeilbarlig metode. For det første oppgir for eksempel NAOB eksplisitt at den øverste betydninga er den eldste. Dette innebærer at den øverste betydninga ikke nødvendigvis er den mest etablerte. Dette er nok også tilfelle i visse artikler i BOB. For det andre er det tenkelig at et ledd har flere betydninger som er etablerte i en slik grad at det blir kunstig å si at sammensetninga bare har høy motivasjon når den øverste betydninga i ordbøker er aktiv i sammensetninga. Dessuten kan kontekst bidra til at ulike betydninger framstår som motiverte.

Videre er det prinsipielt problematisk å dele opp et ords denoteringsevne i forskjellige diskrete betydninger. I mange tilfeller er det rettere å snakke om et betydningskontinuum eller et betydningsfelt hvor ordet kan brukes for å denotere mange tilgrensende eller liknende fenomen, men hvor man i ordbøker av praktiske grunner representerer denne kapasiteten via ulike distinkte betydningsnivå (jf. for eksempel *bleik* i BOB).

Videre er modellen antakelig mer tidkrevende å bruke enn hva som er hensiktsmessig i

leksikografisk sammenheng. Spesielt er faktoren skjematisk produktivitet noe omstendelig å regne ut, i hvert fall når *skjema* svarer til den semantiske relasjonen mellom for- og etterledd. Her må man først fastslå hva slags relasjon mellom for- og etterledd det er i en aktuell sammensetning, gitt en bestemt betydning, for så å undersøke om denne relasjonen er produktiv for det aktuelle for- og etterleddet i andre sammensetninger. Et første problem er at det i mange tilfeller er langt fra innlysende hvordan man skal karakterisere relasjonen mellom sammensetningsledda. Delstudie 1 følger praksisen til Borque (2014); Pepper (2020), hvor ulike relasjoner signaliseres av distinkte parafraseringer. Et problem med denne praksisen er at de mest anomaliserte sammensetningene typisk ikke passer helt overens med en bestemt parafrasering. Sikter for eksempel *tankekors* til et kors **av** eller **i** tanker? Her fins det få argumenter for hvorfor den ene skulle være mer presis enn den andre, og uansett hvilken man velger, er det vanskelig å se for seg at den svarer til et skjema som språkbrukere anvender i fortolkninga av *tankekors*.

I tillegg blir det fort tidkrevende å skulle måle produktiviteten til det skjemaet man velger. Metoden som benyttes i delstudie 1, er å studere de fem mest frekvente sammensetningene som har felles forledd eller etterledd med sammensetninga, og så vurdere om et flertall av disse kan parafraseres på samme måte som sammensetninga som er under vurdering. Altså støter man på den ovennevnte problemstillinga med å relasjonsbestemme sammensetninger hele ti ganger for hver enkelt sammensetning man skal gjennomsiktighetsvurdere. Denne tidsbruken oppveies trolig ikke av gevinsten ved å beholde skjematisk produktivitet som en faktor i modellen. Faktoren bør derfor fjernes eller operasjonaliseres på annet vis.

En faktor som ikke synes å være essensiell for modellen, er intern disambiguering. Svanlund (2002) nevner det som en mulighet at ledda i en sammensetning kan disambiguere hverandre. Selv om dette later til å stemme i visse tilfeller, indikerer resultata fra anomaliseringsmodellen at dette er en perifer egenskap i sammensetningspopulasjonen. Derfor blir det uholdbart si at fravær av intern disambiguering er en kilde til anomali.

Om man – som det fins gode grunner til – fjerner variablene skjematisk produktivitet og intern disambiguering fra modellen, blir de resterende variablene, nemlig *motivasjonsandel*, *motivasjonsgrad* og *skjematiske forbilder*, desto mer sentrale. Førstnevnte er operasjonalisert som en binær variabel ved at sammensetninger som kan parafraseres noenlunde vellykka uten å bruke andre innholdsord enn dem som inngår i sammensetninga, får uttelling for gjennomsiktighet, mens de øvrige ikke får det. En styrke ved denne operasjonaliseringa er at den er rask å bruke. En svakhet er at det ikke fins noen objektiv målestokk for hva som utgjør en «noenlunde vellykka parafrasering». For eksempel kan dette være kinkig å avgjøre for sammensetninger som har metaforiske eller metonymiske betydningsforskyvninger i ledd eller i helheten. Kan for eksempel *tankelesing* parafraseres med 'lesing av tanker' og *svartsinn* med 'svart sinn'? I artikkelen er svaret

ja siden man kan finne autentisk bruk av frasene 'å ha et svart sinn' og 'å lese andres tanker'. Likevel virker det urimelig at *tankelesing* har samme motivasjonssandel som for eksempel *svartlakkere*. Her virker det uheldig at faktoren *motivasjonsandel* har fått en binær operasjonalisering i delstudie 1, særlig siden motivasjonsandel helst bør ses som et graduelt fenomen, på lik linje med motivasjonsgrad. Det er ikke slik at en sammensetnings etablerte betydning enten blir fyllestgjørende dekka av sammensetningsleddas betydninger eller ikke. Snarere kan man si at helhetsbetydninga i varierende grad har tilleggselementer som ikke er eksplisitt uttrykt i sammensetningas form. Det er imidlertid ikke innlysende hvordan man skal operasjonalisere denne gradualiteten, da det ville fordra at man delte opp sammensetningas betydning i atskilte, kvantifiserbare elementer.

Selv om det totalt sett er lite sannsynlig at modellen i delstudie 1 utgjør noen gullstandard for hvordan man bør beregne semantisk anomaliseringsgrad i leksikografisk sammenheng, gir studien verdifull informasjon om relevansen og nytteverdien til de ulike faktorene. Dette er et bidrag til den leksikologiske kunnskapen om sammensetninger generelt og den leksikografiske behandlinga spesielt.

Avslutningsvis må det presiseres at nytten av en anomaliseringsmodell er avhengig av hvor sentral faktoren semantisk anomalisering er når man justerer for andre utvalgskriterier. For eksempel kan det tenkes at korrelasjonen mellom de mest anomaliserte sammensetningene og de mest etablerte sammensetningene er såpass sterk at semantisk anomali blir et marginalt utvalgskriterium når det justeres for de distribusjonsmessige egenskapene til ulike sammensetninger. Dette spørsmålet vil bli videre diskutert i delkapittel 4.3 og i neste kapittel.

4.2 Delstudie 2

Delstudie 2 utgjøres av en vitenskapelig artikkel ved navn «Assessing word commonness – Adding dispersion to frequency», som ble publisert i *International Journal of Corpus Linguistics* i 2022.

4.2.1 Kontekst og bakgrunn

I delkapittel 2.6.3 påpekes det at korpusfrekvens har en betydelig variabilitet på tvers av korpus og korpusdeler, hvilket gjør det til et lite reliabelt korpusmål. Til tross for dette benyttes korpusfrekvens jevnlig som eneste indikator på utbredelse i korpus (og dermed utbredelse i usus). I løpet av de siste cirka femten årene har det heldigvis blitt gjort mye forskning på hva korpusfrekvens kan og ikke kan fortelle oss aleine, og hvilke målemetoder som kan supplere eller erstatte korpusfrekvens som mål på utbredelse i usus.

En forsker som har gått i bresjen for å gjøre korpusmetodikken mer reliabel og valid, og dermed bedre i tråd med statistiske konvensjoner og rettesnorer innenfor vitenskaper som på mange måter er mer empirisk modne enn lingvistikk, er Stefan Gries (andre nevneverdige bidragsytere er omtalt i delkapittel 2.6).

Gries argumenterer for at måling av korpusspredning gir mer stabile og dermed mer reliable innsikter om korpusfenomeners utbredelse i usus (Gries, 2008, 2010). Det er imidlertid ikke åpenbart hvordan man best måler spredning, og Gries vier følgelig mye tid til å evaluere og sammenlikne en stor mengde spredningsmål. Gjennom en håndfull ulike analyser finner han at de 17 spredningsmålene han tar for seg, fordeler seg i fire distinkte klynger med hensyn til hvilke resultater de gir i møte med ulike korpusdistribusjoner (Gries, 2010).

Delstudie 2 skriver seg inn i rekken av spredningsstudier. Den tar mål av seg til å evaluere reliabiliteten og validiteten til korpusfrekvens og fire distinkte spredningsmål hva gjelder å gjenspeile hvilken utbredelse sammensetninger har i usus. Til dette formålet benyttes 273 sammensetninger (omtalt i underkapittel 3.2.2) og en kryssvalideringsprosedyre (se underkapittel 3.1.4).

I delstudie 2 brukes termen *commonness* 'vanlighet' som term for det som ulike mål på korpusutbredelse er ment å predikere. Her i kappa anvendes i stedet termen *etableringsgrad*.

4.2.2 Funn og diskusjon

Delstudie 2 støter også på problemet med at vi ikke har tilgang til all usus, og derfor ikke kan gjøre noen endelige og ubestridelige evalueringer av de fem målenes prediksjoner. En viss evaluering gjøres likevel i delstudie 3 (se delkapittel 4.3 nedenfor). Derimot kan man se på korrelasjonen mellom de fem målene og hvordan deres evne til å gjenspeile korpuseksterne distribusjoner avhenger av distribusjonen i korpuset.

Et viktig funn fra korrelasjonsanalysen i delstudie 2 er at frekvensestimater fra et korpus er mer i tråd med frekvensen i korpuseksternt materiale for sammensetninger som har jevn spredning, altså som har en jevn distribusjon på tvers av korpuset. Det følger da logisk at korpusfrekvenser gir et bedre bilde på frekvens i usus dersom korpusfrekvensverdien stammer fra en jevn distribusjon, altså at korpusfrekvensens validitet styrkes når distribusjonen er jevn. Dette gir god mening. I visse tilfeller kan for eksempel en sammensetning ha høy korpusfrekvens som følge av at den forekommer mange ganger i én tekst. At den gitte teksten inngår i korpuset, er dermed en betingelse for at sammensetninga skal framstå som frekvent. Om teksten ikke var med, ville korpusfrekvensen til sammensetninga vært betydelig lavere. Derfor blir det stort avvik mellom korpusfrekvensene til

den aktuelle sammensetninga når forskjellig materiale blir sammenlikna.

Det er mye mindre sannsynlig at distribusjoner med jevn korpusspredning oppstår tilfeldig, enn at distribusjoner med høy korpusfrekvens gjør det. En distribusjon med jevn spredning har noenlunde lik korpusfrekvens i korpusdelene. Dette viser at forekomstene har en noenlunde regelmessig forekomst. Holdt sammen med høy frekvens vil denne egenskapen indikere høy etableringsgrad. En sammensetning som *bakgrunn* forekommer for eksempel med en slik regelmessighet at vi med en viss sikkerhet kan forvente at frekvensen forblir den samme dersom materialet dobles. Det er nettopp dette jevn spredning forteller oss, at ordet forekommer regelmessig nok til å følge størrelsesfluktuasjoner i datasettet. På den annen side fins det svært tematisk bundne ord med lav spredning som *vandrefugl*, som kun forekommer i svært spesifikke tekster. Om man dobler utvalget, vil det med all sannsynlighet endre frekvensen til *vandrefugl*.

Spredningstall gir imidlertid ingen informasjon om størrelsesordenen til distribusjonen, derfor bør de ideelt suppleres av korpusfrekvens. For eksempel skjelner ikke spredningsmålene mellom hyppige og sjeldne kontekstspesifikke varianter. Om vi utvider eksempelet ovenfor, vil vi ikke av spredningsmål kunne avlese at det er langt flere personer med doktorgrad i medisin enn i lingvistikk, da begge typer er relativt sjeldne fenomener som fordeler seg ujevnt om man tar hele befolkninga i betraktning. Det samme gjelder andre veien. To fenomen som er jevnt fordelt, kan utmerket godt variere mye i størrelsesorden. For eksempel kan vi tenke oss at antall sjøer og antall trær har en forholdsvis jevn spredning på tvers av norske fylker, men det er utvilsomt slik at de tilhører ulike størrelsesordener. Lik spredning har for eksempel også sammensetningene *tankevirksomhet* og *årrekke* i LBK, men sistnevnte har vesentlig større utbredelse totalt sett da den har om lag fem ganger så mange forekomster som førstnevnte.

Totalt sett illustrerer delstudie 2 at spredning er en viktig faktor i den kvantitative vurderinga av ordbokkandidater, både fordi jevn spredning i seg selv sier noe om regelmessigheten til ulike ord, og fordi jevn spredning validerer frekvensberegninger. Analogt med hvordan mål for sentraltendens, som for eksempel gjennomsnitt, alltid bør ledsages av mål for spredning, som for eksempel standardavvik, bør frekvensmål alltid ledsages av spredningsmål. Med frekvensmåls lave reliabilitet i mente bør det være en fast korpuslingvistisk konvensjon å understøtte frekvensmålinger med målinger av spredning.

4.3 Delstudie 3

Delstudie 3 utgjøres av en vitenskapelig artikkel ved navn «Wheat or chaff? A compound selection model based on look-up data», som ble publisert i *International Journal of Lexicography* i 2023.

4.3.1 Kontekst og bakgrunn

Visse ord, fraser og ordelementer i språkteigen er såpass sentrale at det ville være en forsømmelse å utelate dem. Men utover disse ufravikelige oppslagsorda fins det et tilnærmet uendelig antall ordbokkandidater som ikke må, men kan tas med. Spørsmålet er da hvilke av disse kandidatene ordbokbrukerne har mest behov for. Dette brukerperspektivet er langt på vei utgangspunktet for delstudie 3. Selv om søkestatistikk kan brukes for å evaluere i hvilken grad ordboka dekker brukernes behov og interesser, kan det være mange ord som rettmessig tilhører den språklige innmarken, men som ikke nødvendigvis har stor søkeinteresse av den grunn (se delkapittel 2.1). Det er altså ikke noe mål at det skal være et en-til-en-forhold mellom ordboka og søkestatistikken.

Som nevnt flere ganger i denne avhandlinga har vi ingen direkte tilgang på all usus. Vi har derfor ingen garanti for at variablene vi bruker for å rangere ord etter etableringsgrad, gir en presis avspeiling av usus. Leksikografer veit derfor ikke helt sikkert om den lemmalista av sammensatte ord de ender opp med, utgjøres av de n mest etablerte sammensatte ordformene i usus. Men om vi dreier litt på perspektivet og reflekterer over når ordbokoppføring er nyttig, trenger kanskje ikke lemmalista i en ordbok å gi en perfekt gjenspeiling av sammensetningspopulasjonen (altså usus), så lenge den tjener en annen populasjon, nemlig ordbokbrukerne. Grunnen til at korpusutbredelse og anomaliseringsgrad er viktige variabler for leksikografer, er ikke først og fremst at de hjelper oss å finne de mest etablerte sammensetningene, men at de hjelper oss å finne de sammensetningene vi tror brukerne er mest interesserte i.

Populasjonen med ordbokbrukere (heretter *brukerne*) har vi en viss tilgang på. Som omtalt i underkapittel 3.2.3 genereres det for BOB og NOB statistikk av søkene brukerne foretar. Denne statistikken gir et uvurderlig innblikk i brukernes behov og interesser. Som nevnt under omtalen av både delstudie 1 og delstudie 2 kan vi ikke evaluere anomaliseringsmodeller eller spredningsmål ved hjelp av usus, men vi kan til en viss grad evaluere dem med hensyn til hvor treffsikre prediksjoner de gir av brukernes søkeatferd.

Flere tidligere studier har undersøkt korrelasjonen mellom søkeatferd og relevante leksikografiske variabler av både kvalitativ og kvantitativ type (se for eksempel Bäckerud, Nilsson og Sköldberg (2020); Müller-Spitzer, Wolfer og Koplenig (2015); Schryver, Joffe, Joffe og Hillewaert (2006); Trap-Jensen, Lorentzen og Sørensen (2014); Wolfer, Koplenig, Meyer og Müller-Spitzer (2014)). Så vidt jeg kan se, undersøkes imidlertid kun én kvantitativ prediktor og én kvantitativ responsvariabel i disse studiene, og ikke uventa er det korpusfrekvens og oppslagsfrekvens som anvendes (eller i visse tilfeller også antall forekomster i korpus og antall oppslag). Med andre ord lider samtlige av disse studiene av det problemet som Gries (2022a) beskriver, at det er lett å finne signifikante korrelasjoner mellom korpusfrekvens og en responsvariabel når nullhypotesen er at korpusfrekvens

ikke spiller noen rolle, og det dessuten ikke kontrolleres for andre kvantitative variabler. Korpusfrekvens absorberer såleis all kvantitativ informasjon man har om observasjonene sine. I tillegg er også responsvariabelen i disse studiene frekvensbasert, hvilket vil si at også den er utsatt for høy variabilitet slik som korpusfrekvens (se omtale i delkapittel 2.6).

I delstudie 3 undersøkes derfor sammenhengen mellom korpusfrekvens, spredning og en rekke kvalitative variabler (deriblant motivasjonsgrad) på den ene sida og oppslagsregularitet på den andre (se omtale i underkapittel 3.2.3). Studien er imidlertid mindre opptatt av korrelasjon enn de ovennevnte studiene. Grunnen til dette er at veldig mange ordformer (også sammensetninger) aldri befinner seg på vippepunktet mellom å bli oppført eller utelatt i leksikografisk sammenheng. Blant disse inngår ordformene som befinner seg i det øverste frekvenssjiktet, og som i henhold til de ovennevnte studiene også har svært høy oppslagsfrekvens. Det at det sannsynligvis er en nær korrelasjon mellom korpusutbredelse og oppslagsregularitet i det høyeste frekvenssjiktet, kan ha en stor innvirkning på korrelasjonskoeffisienten mellom disse to variablene. En korrelasjonsanalyse kan derfor gi feilaktige inntrykk av sammenhengen mellom disse variablene i lavere frekvenssjikt. Dette er uheldig ettersom det er nettopp i lavere frekvenssjikt at leksikografer trenger utvalgsvariabler for å identifisere de mest relevante ordbokkandidatene.

Med ovennevnte argument i mente kunne man helt enkelt fjerna det øverste frekvenssjiktet fra korrelasjonsanalysen. Dette har imidlertid den åpenbare svakheten at det ville vært vanskelig å finne et godt grunnlag for å sette den optimale grenseverdien mellom øvre og nedre frekvenssjikt. Uansett hvor høy grenseverdi man setter, vil den kunne ha en arbitrær påvirkning på de etterfølgende analysene. Videre har leksikografer bruk for å vite ikke bare hvilke utvalgsvariabler de skal bruke, altså hvilke variabler som er korrelert med søkeinteresse, men også hvordan de skal bruke dem. Korrelasjonkoeffisienter gir derimot få svar på hvordan en variabel reint praktisk kan anvendes for å skille ut de mest relevante sammensetningene.

I delstudie 3 legges det derfor mest vekt på en metode som undersøker hvilke utvalgsvariabler og deretter hvilke verdier av disse variablene som hjelper en å skille sammensetninger som later til å være uinteressante for brukerne, fra sammensetninger som later til å være interessante. Til formålet benyttes to ikke-parametriske metoder kalt inferenstrær og randomiserte skoger (se detaljert omtale i 3.1.5). Styrken med disse metodene er at de kaster lys over hvilke prediktorer som i størst grad predikerer systematiske forskjeller i oppslagsregularitet, hvilke nivåer av disse prediktorene som maksimerer forskjell i oppslagsregularitet dersom man deler observasjonene i to, og hvordan flere prediktorer sammen kan hjelpe en å skille ut grupperinger av sammensetninger med lav eller høy oppslagsregularitet. For eksempel kan det tenkes at sammensetninger med høye spredningsverdier har gjennomgående høy oppslagsregularitet, mens det blant de resterende

sammensetningene er stor forskjell på sammensetninger med høy og lav motivasjonsgrad, hvor sistnevnte har sammenheng med høy oppslagsregularitet. Denne metoden gir altså noen svar på hvilke prediktorer som hjelper en å finne de mest relevante ordbokkandidatene, og videre under hvilke omstendigheter prediktorene har denne kapasiteten.

Søkestatistikken utnyttes dessuten til å evaluere sammensetningsutvalget i BOB. Til dette formålet undersøkes det i hvilken grad BOB eventuelt mangler sammensetninger med høy oppslagsregularitet, kalt *lakuner*, og i hvilken grad BOB har med sammensetninger uten dokumentert søkeinteresse, kalt *ubesøkte oppslagsord*.

Søkestatistikken som anvendes i delstudie 3, gir imidlertid ingen informasjon om brukernes intensjon med å søke opp et visst ord. For eksempel er det tenkelig at visse sammensetninger primært søkes opp fordi brukeren vil dobbeltsjekke staving, bøyning eller hvorvidt de skrives med fuge-s, mens andre søkes opp for å undersøke typisk bruk, betydning eller om sammensetninga «fins». Det er lett å tenke seg at mindre sikre språkbrukere, som for eksempel innlærere, vil finne det betryggende at sammensetninga de ønsker å bruke, står oppført i ordbøker. Siden brukere konsulterer ordbøker av ulike grunner (se f.eks. Pilke (2008)), ville det vært gunstig å ha mer detaljert informasjon om søkeatferden. Med sånne data kunne man for eksempel undersøkt om det var en korrelasjon mellom søk på ugjennomsiktige sammensetninger og søk med formål om å undersøke betydning og typisk bruk.

4.3.2 Funn og diskusjon

Funna i delstudie 3 viser at den eksisterende lemmalista i BOB i forholdsvis stor grad fanger de mest relevante sammensetningene med hensyn til hva brukerne ifølge søkestatistikken later til å være interesserte i. BOB har en tilsynelatende lav andel lakuner og ubesøkte oppslagsord. Treffsikkerheten til BOB er dessuten bedre enn prediksjonsevnen til modellen som jeg har utvikla basert på resultater fra inferenstrær og randomiserte skoger. Dette tyder på at leksikografene som redigerer BOB, allerede benytter relativt treffsikre metoder for å gjøre hensiktsmessige utvalg av sammensetninger.

I delstudie 3 blir det dessuten undersøkt i hvilken grad det er korrelasjon mellom BOB-status og blant annet korpusfrekvens, -spredning og motivasjonsgrad. Her kommer det fram at de aller fleste sammensetninger i de høyeste frekvens- og spredningssjikta er ordbokført. Samtidig er det mange sammensetninger med lav frekvens og/eller ujevn spredning som også er ordbokført. Dette tyder på at sammensetninger i liten grad blir diskvalifisert fra ordbokføring på bakgrunn av variablene korpusfrekvens og -spredning. Motivasjonsgrad på sin side har en veldig usikker og i beste fall marginal assosiasjon med ordbokstatus. Selv om et flertall av sammensetningene som avviker fra full motivasjon er ordbokført, er det såpass få sammensetninger i denne gruppa at man vanskelig kan

konkludere med noe. Blant maksimalt motiverte sammensetninger framgår det heller ikke noe tydelig mønster med hensyn til BOB-status.

På overordna plan viser funna fra delstudie 3 at sammensetninger med høy korpusutbredelse med stor sannsynlighet også har høy oppslagsregularitet. For eksempel blir høyfrekvente og godt spredde sammensetninger som *forutsette*, *fortid*, *foruten*, *forut*, *dømmekraft* og *dødsstraff* jevnlig søkt på i standardordbøkene.

Inferenstreanalysene viser for det første at variasjoner i oppslagsregulariteten til sammensetninger bestående av to nominaler ikke fullgodt lar seg forklare av prediktorene i studien, i hvert fall ikke i samme grad som sammensetninger med minst en ikke-nominal. For sistnevnte gruppe er det betydelig forskjell på oppslagsregulariteten til sammensetninger i øvre og nedre sjikt av spredningsskalaen (som er 0–1), hvor sammensetninger med jevn spredning som forventa har gjennomgående høy oppslagsregularitet. I denne gruppa fins blant annet *akterut*, *bråvåkne*, *dødsdømt* og *forventningsfull*. Blant ikke- og semi-nominale sammensetninger ser det dessuten ut til at motivasjonsgrad er en nyttig prediktor i nedre sjikt av spredningsskalaen. Blant ikke- og seminominale sammensetninger med ujevn spredning, lav motivasjonsgrad og høy oppslagsregularitet finner vi blant annet *ageløs*, *brønnpisser*, *dødball*, *forvei* og *forutsetningsløs*.

For nominale sammensetninger er det mye uforklart variasjon. En kombinasjon av høy frekvens og jevn spredning hjelper en å skille ut en liten gruppe med sammensetninger med høy oppslagsregularitet. For eksempel fins sammensetningene *aksjeselskap*, *brødskive*, *dødsfall*, *ekskjæreste* og *forstørrelsesglass* i denne gruppa. Utover dette er det ingen av prediktorene i studien som aleine eller i kombinasjon filtrerer ut nominale sammensetninger med lav eller høy oppslagsregularitet på en treffsikker måte. Prediktorene i studien kommer altså til kort hva gjelder å gi en fullgod forklaring av hva som styrer brukernes søkeinteresse og -behov.

Totalt sett forteller delstudie 3 oss at det antakelig er svært mange variabler som forklarer brukernes søkeatferd. Trolig vil noe av variasjonen i oppslagsregularitet forbli uforklart, men det bør testes om for eksempel ordlengde, billedlighet, flere korpus, onomasiologisk prominens, ordhistorie, ordopphav og liknende kan forklare mer av søkevariasjonen. Dessuten viser resultata i delstudie 3 oss at ingen av de undersøkte variablene hjelper oss å utelukke søkeinteresse. Selv om en sammensetning har ujevn spredning eller ikke forekommer i korpus i det hele tatt, kan den framleis dukke opp i søkestatistikken. Om man ønsker å kompilere ordbøker som i størst mulig grad fanger søkeinteressen, er det ingen av variablene i studien som aktivt diskvalifiserer en sammensetning for oppslag. Verken ujevn spredning, lav frekvens eller høy motivasjonsgrad gir grunnlag for å utelukke sammensetninger fra ordboka, mens det motsatte, altså jevn spredning, høy frekvens eller lav motivasjonsgrad, kan gi grunnlag for å ta dem med (se videre diskusjon i neste kapittel).

På et eller annet vis kan det virke som leksikografene som har redigert og per dags dato redigerer BOB heller ikke ekskluderer sammensetninger på bakgrunn av variablene som undersøkes i denne studien. Lemmalista til BOB ser nemlig ut til å ha relativt tilfredsstillende dekning av brukernes behov. Selv om det ikke er mulig å spore beslutningsgrunnlaget leksikografene har basert inntaket av de forskjellige sammensetningene på, virker det sannsynlig at leksikografene samla sett har en velutvikla intuisjon for hvilke sammensetninger brukerne slår opp. Trolig er det denne intuisjonen som gjør leksikografene mer treffsikre i sin prediksjon av søkeinteresse enn den multivariate modellen i delstudie 3. Likevel fins det framleis både lakuner og ubesøkte oppslagsord i BOBs lemmaliste innenfor det utvalget som anvendes i delstudie 3. Om antallet lakuner og ubesøkte oppslagsord i det aktuelle utvalget er representativt, vil det totale antallet lakuner og ubesøkte oppslagsord fra A til Å være betydelig, så det er all grunn til å forsøke å øke treffsikkerheten ytterligere. Den nygenererte modellen i delstudie 3 lyktes som nevnt ikke i å oppnå en bedre prediksjonsevne, men det er framleis mange viktige funn i analysene som ligger til grunn for denne modellen, som gir håp om at økt treffsikkerhet kan oppnås gjennom videre forskning. Samtidig må man også ha i mente at det ikke bare er søkestatistikken som skal informere sammensetningsutvalget i allmennordbøker (se diskusjon i underkapittel 4.3.1).

Kapittel 5

Variabler til leksikografisk seleksjon

I dette kapittelet samler jeg trådene fra den teoretiske inngangen og de tre delstudiene som er omtalt ovenfor. Innledningsvis konstateres noen overordna forutsetninger for språk og leksikografi, før jeg diskuterer både etablerte kjensgjerninger og usikre indikasjoner knytta til den leksikografiske utvelgelsen av sammensatte ord. Deretter foreslår jeg en samling variabler som bør inngå i en prosedyre for leksikografisk seleksjon av sammensetninger. Til slutt testes prosedyren på en samling potensielle sammensetninger.

Den initielle problemstillinga til avhandlinga dreier seg om misforholdet mellom en ordboks finitte natur og dens målobjekts dynamiske, uendelige og omskiftelige natur. En språkvarietets leksikalske ressurser er for fleksible og rike til at de i sin helhet lar seg fange mellom to permer, eller for den del i en database. Det sammensatte ordtilfanget til norsk speiler diversiteten og fleksibiliteten i språkbrukeres generelle orddannings- og ordprosesseringsevne. Som Benczes (2006) påpeker, er den menneskelige orddanningsevnen kjennetegna av kreativitet. Selv om jevne språkbrukere har en internalisert evne til å skape regelmessige sammensetninger, har de også kreativiteten til å bryte mønsteret og til å videreføre bruken av relativt anomaliserte sammensetninger. Resultatet blir derfor et tilfang bestående av sammensetninger med varierende etableringsgrad og grad av semantisk og morfologisk anomalisering.

Leksikografiens oppgave blir da for det første å skille etablerte sammensetninger fra uetablerte sammensetninger. En prinsipiell grunn til at ordbøker bør styre unna uetablerte sammensetninger, er at ordbokføring oppfattes som en konvensjonaliserende og normerende kraft, selv om ordboka skulle etterstreve deskriptivitet. Det er i tråd med den deskriptive ambisjonen at orda som velges for leksikografisk beskrivelse, allerede er konvensjonaliserte, slik at ordboka reelt sett fanger konvensjonelle ord og uttrykksmåter framfor å utøve konvensjonaliserende makt på ukonvensjonelle ordformer. For det andre må kandidatene rangeres slik at det er de mest relevante sammensetningene som selekteres for leksikografisk beskrivelse.

5.1 Kjensgjerninger

Under følger en liste over noen sentrale kjensgjerninger som enten har blitt styrka eller fastslått i denne avhandlinga. Kjensgjerning 5 er et nytt funn som i hvert fall i norsk sammenheng først ble påpekt i delstudie 3 i denne avhandlinga.

1. Sammensetninger har ulik etableringsgrad.
2. Sammensetninger har ulik gjennomsiktighetsgrad.
3. Utvelgelse av sammensetninger fordrer en multivariat tilnærming.
4. Det er vesentlig mer tidkrevende å gjøre detaljerte beregninger av semantisk anomaliseringsgrad ved sammensetninger enn kvantitative beregninger av frekvens og spredning.
5. Også lavfrekvente, ujevnt spredde og fullt motiverte sammensetninger blir søkt etter.

At sammensetninger har varierende etableringsgrad, er omhyggelig etablert gjennom kapittel 2, der jeg forklarer hvordan sammensetninger kan variere i konvensjonaliserings- og innprentingsgrad. Dette henger sammen med hvordan sammensetninger benyttes til ulike formål (se 2.4), hvor et av formåla er å gi et begrep en stabil betegnelse. Sammensetninger som benyttes til dette formålet, vil med stor sannsynlighet få høyere etableringsgrad enn sammensetninger som benyttes som midlertidige adhoc-uttrykk for et konkret objekt i en avgrensa kontekst, f.eks. *votteskuff*.

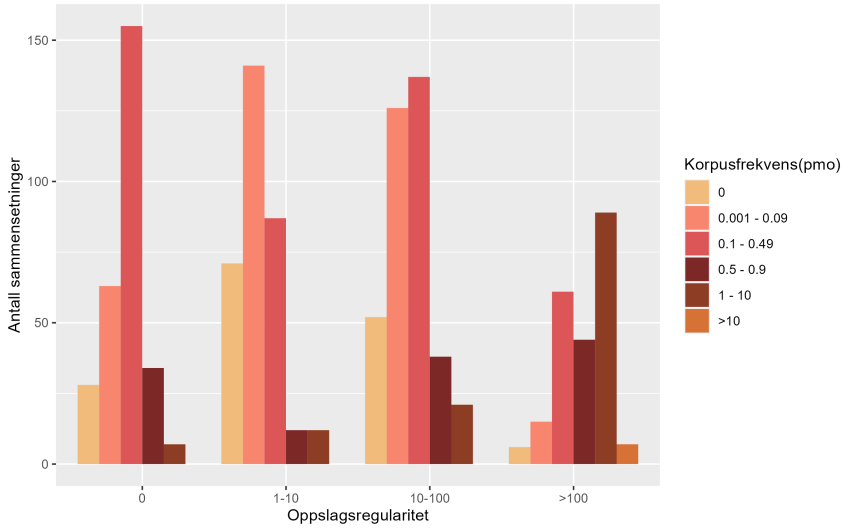
At sammensetninger har varierende gjennomsiktighetsgrad, kan utledes ved å studere mer eller mindre etablerte sammensetningers semantiske og morfologiske anomaliseringsgrad (se detaljert omtale i delkapittel 2.5). Variasjoner i anomaliseringsgrad beskrives for øvrig i avhandlingene til Bakken (1998) og Eik (2019) (under termen *leksikalisering*).

At utvelgelse av sammensetninger krever en multivariat tilnærming, er en nødvendig konsekvens av at etableringsgraden påvirkes av både kvantitative og kvalitative variabler, slik det framgår i delkapittel 2.4. Dessuten påviser blant annet delstudie 2 at kvantitative beregninger av utbredelse bør være multivariate, og delstudie 1 at kvalitative beregninger av gjennomsiktighetsgrad også bør være multivariate.

At det er tidkrevende å gjøre detaljerte og multivariate anomaliseringsberegninger, gir seg nærmest selv, men det er verdt å trekke fram for å understreke at semantisk anomalisering beror på mange faktorer, hvorav enkelte er særlig tidkrevende å beregne og måle for et konkret sammensatt ord. Det fins en stor og mangefasettert forskningslitteratur

knytta til spørsmålet om semantisk anomalisering, hvorav noen sentrale eksempler refereres i delkapittel 2.5. Om man skal bryte innsiktene fra denne litteraturen ned i aktuelle anomaliseringsvariabler, vil man ende opp med langt flere variabler enn hva som er tjenlig for en leksikografisk prosedyre. I delstudie 1 ble det forsøkt en multivariat tilnærming til semantisk anomalisering, mens motivasjonsgrad ble brukt som eneste indikasjon på anomalisering i delstudie 3. Resultata fra disse noe heuristiske tilnærmingene kan tyde på at anomalisering enten bør operasjonaliseres mer treffsikkert, eller at det aleine er et marginalt mål på ordbokrelevans. Uansett tyder resultata fra delstudie 3 på at anomaliseringsgrad først og fremst bør anvendes for å skille ut relevante sammensetninger i det lavfrekvente og ujevnt spredde sjiktet.

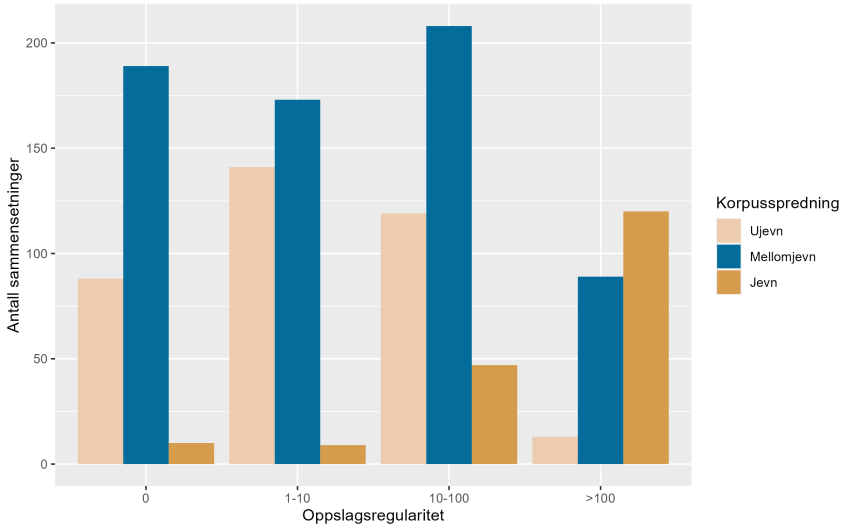
Kjensgjerning 5 framgår av figur 5.1, 5.2 og 5.3. I samtlige av disse figurene fordeles sammensetninger innenfor ulike sjikt av oppslagsregularitet (se underkapittel 3.2.3 for definisjon) langs x-aksen. I tillegg angis ulike sjikt av korpusforekomster med ulike farger, mens y-aksen viser antall sammensetninger innenfor hver kategori. Figurene indikerer med andre ord korrelasjonen mellom henholdsvis korpusfrekvens, korpusspredning og motivasjonsgrad på den ene siden og oppslagsregularitet på den andre.



Figur 5.1: Antall forekomster fordelt etter oppslagsregularitet fra delstudie 3

I figur 5.1 kan man helt til høyre se at det fins en liten andel sammensetninger med null eller få korpusforekomster som har over 100 i oppslagsregularitet. Noen eksempler herfra er *ageløs*, *brønnpisser*, *dødgang*, *ekkokammer* og *forutbestemmelse*. Av disse kan *ekkokammer* forklares med at bruken av denne sammensetninga ifølge Nasjonalbibliotekets n-gramsøk har økt ekspansivt etter tidsperioden korpuset dekker, mens de resterende har hatt jevn eller til og med dalende bruk de siste 100 årene. Uansett viser denne gruppa

at korpusfrekvens overser visse sammensetninger som brukere jevnlig søker informasjon om.

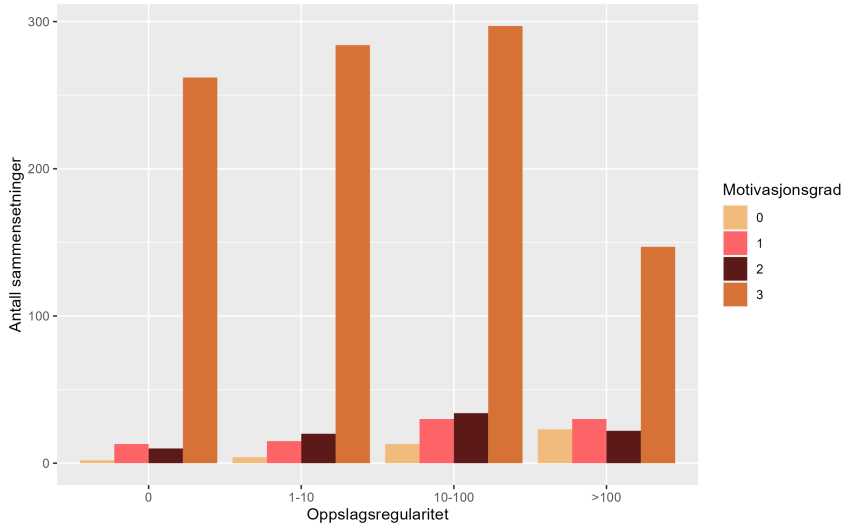


Figur 5.2: Spredningsverdier fordelt etter oppslagsregularitet fra delstudie 3

Figur 5.2 viser at en liten gruppe sammensetninger med ujevn spredning har over 100 i oppslagsregularitet. Det er en viss overlapp mellom denne gruppa og tilsvarende gruppe knytta til korpusfrekvens som omtales i forrige avsnitt. Blant annet inngår for eksempel *ageløs*, *brønnpisser*, *dødgang* og *ekkokammer* i begge. To sammensetninger med høy søkeinteresse som har ujevn spredning og en viss frekvens, er *aksenttegn* og *dødsrate*. Totalt sett indikerer figur 5.2 at også en del sammensetninger med ujevn korpusspredning blir søkt mye etter.

I figur 5.3 kan man dessuten stadfeste at en ganske stor gruppe sammensetninger med maksimal motivasjonsgrad har over 100 i oppslagsregularitet. Blant disse fins for eksempel *aksjeselskap*, *brødskive*, *dørhåndtak*, *ekskone* og *fortellerstemme*. I motsatt ende finner vi to sammensetninger med minimal motivasjonsgrad og null i søkeinteresse, nemlig *dvergstein* og *dødkjøtt*. Man kan se på de to stolpene lengst til venstre i hver stolpeklynge at antallet sammensetninger med lav motivasjon øker i takt med oppslagsregulariteten, men det er uansett klart at også sammensetninger med maksimal motivasjonsgrad blir søkt etter ganske jevnlig.

I dette delkapittelet har jeg oppsummert noen kjensgjerninger som har blitt etablert forut for og underveis i denne avhandlinga. I neste delkapittel følger noen indikasjoner som framgår av delstudiene, men som behøver ytterligere belegg før vi eventuelt kan betrakte dem som kjensgjerninger.



Figur 5.3: Motivasjonsgradsverdier fordelt etter oppslagsregularitet fra delstudie 3

5.2 Indikasjoner

De følgende indikasjonene regner jeg som usikre fordi det foreløpig ikke fins tilstrekkelig empirisk belegg for dem til at de kan opphøyes til kjensgjerninger. Det er for eksempel tenkelig at ny empiri kan peke i en annen retning, og at påstandene dermed må modereres. Indikasjonene er likevel sterke nok til at de vil tjene som kunnskapsgrunnlag for prosedyren som foreslås seinere i dette kapittelet.

At korpusfrekvens, korpusspredning og motivasjonsgrad er uegna som eksklusjonskriterier, kan utledes av kjensgjerning 5 og figur 5.1, 5.2 og 5.3. Vi kan med sikkerhet slå fast at selv sammensetninger med pessimale verdier, det vil si de verdiene som indikerer lavest etableringsgrad, for samtlige av disse tre variablene, blir søkt etter en hel del av brukerne av standardordbøkene. Helt konkret fins det 47 slike sammensetninger i et datautvalg på 1206 sammensetninger i delstudie 3. Dette forholdet indikerer at man vil gå glipp av viktige sammensetninger dersom man benytter en eller flere av de tre variablene som eksklusjonskriterier. Om vi hypotetisk utelukka alle sammensetninger i LBK med korpusfrekvens under 0,1 per million ord, ville det mest sannsynlig ført til at tusenvis av relevante sammensetninger ble oversett. Siden korpus etter alt å dømme ikke fanger opp samtlige relevante sammensetninger, fungerer korpusfrekvens og korpusspredning dårlig som kriterier for å ekskludere ordbokkandidater. Hva gjelder motivasjonsgrad, viser figur 5.3 at høy motivasjonsgrad er dominerende innenfor samtlige oppslagsregularitetssjikt. Det ville med andre ord bli mange lakuner dersom man ekskluderte sammensetninger basert på at anomaliseringsgraden indikerte at de var maksimalt gjennomsiktige.

At sammensetningsutvalg ideelt bør kompileres ved hjelp av en rekke forskjellige inklusjonsgrunnlag, framfor et sett med rigide kriterier som alle sammensetninger må oppfylle, kan utledes av delstudie 3 og det som oppsummeres i indikasjon 1. Et kriterium eller et sett kriterier benyttes for å enten finne sammensetninger å inkludere eller ekskludere basert på en eller flere betingelser, for eksempel en grenseverdi av korpusfrekvens og -spredning. Fordelen med å basere utvelgelse av sammensetninger på inklusjonskriterier framfor eksklusjonskriterier er at vi anerkjenner at ordboksøk kan ha forskjellige motiver som ikke er betinga av hverandre. Ordbokbrukere som har et konkret ord de ønsker å slå opp, er likegyldige med hvor ofte dette ordet forekommer i usus, eller hvor høy anomaliseringsgrad det har. Korpusutbredelse og anomaliseringsgrad kan naturligvis øke sannsynligheten for at en bruker støter på eller stusser over et ord og dermed slår det opp i en ordbok, men de er langt ifra betingelser for at dette skal skje. Snarere er det slik at utbredelse i usus eller anomaliseringsgrad er faktorer som, uavhengig av hverandre, kan motivere eller sannsynliggjøre søkeinteresse. De representerer derfor to separate grunnlag for å innlemme en sammensetning i ei ordbok. Som funna i delstudie 3 viser, kan evnen til å fange de fleste sammensetningene som fra et brukerperspektiv framstår som relevante, forbedres ved å identifisere flere faktorer som motiverer ordbokføring; mer om dette i delkapittel 5.3.

At det er forskjell på variasjonen blant nominale sammensetninger på den ene siden og ikke- og seminominale på den andre, framgår av funna i delstudie 3. I denne studien brukes inferenstrær og randomiserte skoger (se underkapittel 3.1.5) til å identifisere en utvelgelsesmodell for sammensetninger basert på søkestatistikk. Denne modellen og det eksisterende lemmautvalget i BOB ble i tillegg evaluert med henblikk på søkeinteresse (se detaljert beskrivelse i delkapittel 4.3). Tabell 5.1 viser andelen lakuner og ubesøkte oppslag blant nominale og ikke- og seminominale sammensetninger i henholdsvis BOB og den nye utvelgelsesmodellen som utvikles i delstudie 3. Der lakuner kan være uheldige siden de representerer ord som brukerne søker etter, men ikke får informasjon om, kan ubesøkte artikler ha en viss nytteverdi gjennom at de for eksempel kan bli henvist til fra andre artikler i ordboka, eller at de inngår i ordsystemer. I BOB er lakuneandelen nesten helt lik for nominaler på den ene siden, og ikke- og seminominaler på den andre, mens det er stor forskjell på disse gruppene i utvelgelsesmodellen. I tillegg er andelen ubesøkte oppslag større for nominaler i begge modeller. Disse talla indikerer at både BOB og utvelgelsesmodellen gjør et mer ufiltrert inntak av nominale sammensetninger enn av ikke- og seminominaler, uten at dette gjør lakuneandelen til nominalene lavere. Siden andelen ubesøkte oppslag for ikke- og seminominaler er forsvinnende liten, er det grunn til å tro at den eksisterende ordbokredaksjonen til BOB og i enda større grad utvelgelsesmodellen har høy spesifisitet med hensyn til utvalg av ikke- og seminominale sammensetninger, i alle fall sett til søkestatistikken.

Modell	Delgruppe	% lakuner	% ubesøkte oppslag
BOB	Nominaler	0,17	0,16
	Ikke- og seminominaler	0,18	0,02
Utvelgesesmodellen	Nominaler	0,23	0,12
	Ikke- og seminominaler	0,08	0,05

Tabell 5.1: Oversikt over andel lakuner og ubesøkte oppslag for ulike delgrupper

Mens BOB presterer jevnt med hensyn til lakuner i de to delgruppene, gjør utvelgelsesmodellen det vesentlig dårligere på nominaler. Siden antallet nominale sammensetninger er så stort i norsk, vil en lakunerate på 0,23 kunne svare til flere tusen sammensetninger som brukerne søker etter uten å få treff. Med andre ord fins det søkeinteressevariasjon blant nominale sammensetninger som variablene i modellen og i delstudie 3 ikke forklarer, og vi behøver flere variabler for å kunne fange lakunene uten at det genererer for mange ubesøkte oppslag.

I dette delkapittelet har jeg presentert og diskutert følgende indikasjoner fra delstudiene:

1. Verken korpusfrekvens, korpusspredning eller motivasjonsgrad er gode eksklusjonskriterier for sammensetninger.
2. Det optimale sammensetningsutvalget kompileres ved hjelp av et sett inklusjonskriterier.
3. Det kreves flere variabler for å forklare variasjonen i søkestatistikk blant nominale sammensetninger enn blant ikke- og seminominale sammensetninger.

I det følgende vil jeg med utgangspunkt i kjensgjerningene og indikasjonene ovenfor foreslå en samling variabler for seleksjon av sammensetninger til allmennordbøker og en mulig operasjonalisering av disse.

5.3 Hensiktsmessige variabler

Med de ovenstående kjensgjerningene og indikasjonene i mente kan det for det første påpekes at det vil være gunstig med flere empiriske undersøkelser som forsøker å forklare variasjonen i søkeinteresse særlig blant nominale sammensetninger. For det andre vil det være gunstig med empiriske undersøkelser som tar sikte på å identifisere flere relevante variabler knytta til ordbokaktualitet. For det tredje vil det være gunstig med mer forskning på hvordan ulike konstellasjoner av variabler og operasjonaliseringer av disse presterer i møte med brukernes dokumenterbare søkeinteresse (se videre diskusjon av framtidig forskning i delkapittel 6.2).

I det følgende vil jeg først diskutere de viktigste ingrediensene i et gunstig sammensetningsutvalg, før jeg framhever en gruppe variabler som kan benyttes for å skjelne mellom gode og mindre gode ordbokkandidater. Deretter vil jeg komme med et forslag til en leksikografisk prosedyre for å identifisere gode ordbokkandidater med utgangspunkt i disse variablene. Prosedyren vil være mynta på norsk, og dermed den språklig virkeligheten og ressursene som fins i en norsk kontekst. Den vil imidlertid være uproblematisk å oversette til andre språk og språksamfunn. I beskrivelsen av prosedyren vil jeg ikke vektlegge det faktum at det fins to norske skriftspråk, men heller forutsette at den kan følges uavhengig av hva slags skriftspråk ordboka er ment å dekke.[1]

Schmid (2020) gir en overbevisende og detaljert framstilling av det han kaller *The dynamics of the linguistic system*, nemlig drivkreftene bak språklig tilblivelse og endring. Ifølge Schmid fins det to hoveddrivkrefter, *konvensjonalisering* og *innprenting*, som igjen er knytta til hver sine underordna prosesser. Konvensjonalisering utgjøres av prosessene *usualisering* og *diffusjon*, mens *innprenting* blant annet utgjøres av prosessene *rutinisering* og *skjematisering* (se delkapittel 2.2 og Schmid (2020)) for mer detaljert beskrivelse av disse begrepa). Man kan se disse prosessene som faktorer for etableringsgraden til ord og uttrykk. Hvor etablert for eksempel en sammensetning er i språksamfunnet, avhenger av dens usualisering og diffusjon, mens hvor etablert den er for en gitt språkbruker, avhenger av dens grad av rutinisering og skjematisering. Innprenting og konvensjonalisering har gjensidig påvirkning på hverandre gjennom språkbruk. Konvensjonelle ord og uttrykk blir oftere tatt i bruk og derfor lettere innprenta hos språkbrukere, samtidig som innprenta ord og uttrykk lettere blir tatt i bruk av språkbrukerne og såleis ytterligere konvensjonalisert gjennom diffusjon.

Ordboktypen som den følgende prosedyren er mynta på, kan i henhold til Fjeld og Vikørs (2008, 134–140) ordboktypologi beskrives som hovedsaklig allmennspråklig, synkron, enspråklig, semasiologisk, beskrivende og syntagmatisk (men dette betyr ikke at prosedyren er ubrukelig for ordbøker som har innslag av disse parametrenes motstykker, altså som er delvis spesialspråklige, diakrone, tospråklige, onomasiologiske, normative eller paradigmatiske). Sagt på en bedre måte gjelder prosedyren for ordbøker som har som mandat å beskrive de viktigste orda og uttrykka i det felles, samtidige, bruksspråket, slik som for eksempel standardordbøkene. Disse inneholder primært språklige betegnelser som med en viss sannsynlighet påtreffes i allment tilgjengelige og brukte fora, innenfor en hensiktsmessig tidshorisont som strekker seg fra i dag og et gitt antall tiår bakover. Som nevnt i delkapittel 2.1 er det likevel ikke slik at allmennordbøker kan utelukke alt som kan kalles fortidige eller spesialspråklige ordformer. Ordbokas klassifisering er nok derfor først og fremst en beskrivelse av primærambisjonen til det aktuelle ordbokverket.

[1]For redaksjonen som redigerer både BOB og NOB, vil prosedyren simpelthen måtte bli fulgt to ganger, en gang for hvert skriftspråk. Siden det totalt sett fins større og mer utbygde ressurser for bokmål, vil nok sammensetningsutvalget i NOB til en viss grad også informeres av utvalget i BOB.

En grunnbetingelse for alle ord som tas med i allmennordbøker, er ifølge Fjeld og Vikør (2008, 156) autentisitet. Ordet må ha blitt brukt i en gitt betydning for at ordet og den korresponderende betydninga kan beskrives leksikografisk. I Schmids termer kan man si at sambandet mellom ordet og betydninga må være nokså semasiologisk og onomasiologisk usualisert innenfor et språkmiljø, altså at en viss andel av språkmiljøets medlemmer har en taus overenskomst om betydninga gitt ordet, og at ordet er en aktuell betegnelse for betydninga. I en allmennordbok er i prinsippet alle usualiserte sammensetninger aktuelle og mulige ordbokoppføringer.

Alle medlemmer av et gunstig sammensetningsutvalg bør derfor være usualiserte. Videre består et gunstig sammensetningsutvalg i allmennordbok av de mest etablerte, de mest anomaliserte og de mest ettersøkte sammensetningene. Dersom et sammensetningsutvalg utgjøres av det akkumulerte medlemstilfanget av disse tre undergruppene, sikrer man at de orda som er mest ordinært i bruk, de orda som er mest ugjennomsiktige, og de orda som brukerne har størst behov for å slå opp, er representert i ordboka. Resultatet vil da bli en ordbok som både tjener som et nyttig verktøy for språkbrukerne og som en treffsikker representasjon av ordtilfanget i den språklige varieteten.

For å identifisere de mest etablerte, anomaliserte og ettersøkte sammensetningene i språkteigen, kan dermed følgende samling variabler benyttes:

- Diffusjon
- Anomalisering
- Skjematisering
- Usualiseringsdomene
- Erfaringsbasert innprenting
- Oppmerksomhetsverdi
- Videre orddanning

Diffusjonsgraden avgjøres av spredninga til det usualiserte sambandet mellom en sammensatt ordform og dens betydning på tvers av språksamfunnet. Ord med høy diffusjonsgrad er svært konvensjonelle og derfor velkjente for de aller fleste talerne av et språk. Diffusjonsgraden avtegner seg i språkbruk, og for å oppnå høy diffusjonsgrad må orda brukes på tvers av ulike brukssituasjoner. De må ifølge Schmid (2020, kap. 9) være diffundert på tvers av for eksempel geografi, klasse og stil. I ordboksammenheng er det også relevant med diffusjon på tvers av tid. Siden diffusjonsgraden kan spores i undersøkelser av språkbruk, er korpusberegninger relevante mål på diffusjon. Derfor er både korpusspredning og korpusfrekvens, som det påpekes i delkapittel 5.2, mulige inklusjonskriterier

for sammensatte ord, nettopp fordi de på ulike måter indikerer konvensjonalisering gjennom diffusjon. Ujevn spredning eller lav frekvens i korpus gir imidlertid ingen garanti for at en sammensetning har lav diffusjon i usus. Siden datamengden i usus er mange ganger større en datamengden i korpuset, er det sannsynlig at det fins flere diffunderte sammensetninger i usus enn i korpus. For å identifisere flest mulig av de ususdiffunderte sammensetningene, bør man søke i mer enn ett korpus. Dessuten er diffusjon bare et av flere mulige inklusjonsgrunnlag; altså kan selv sammensetninger med lav diffusjon kvalifisere for oppføring via sine andre egenskaper.

Høy anomaliseringsgrad kan gi grunnlag for oppføring for sammensetninger som nettopp bryter etablerte skjema, semantisk eller morfologisk. *Fedreland* representerer for eksempel en anomali siden det er svært sjelden forleddet i norske sammensetninger står i flertall. *Kråketær* representerer en anomali siden det ikke bare betegner en underkategori av *tær*, men også uleselig håndskrift. Dessuten brukes *kråketær* nesten utelukkende i flertall. Billedligheten i *kråketær* er slett ikke uvanlig, men slik billedlighet representerer likevel noe ikke-skjematisk, som i denne avhandlinga kalles anomali. Det ville vært rart å inkludere sammensetninger som *måketær* eller *spurvetær*, da disse ikke har noen annen betydning enn den maksimalt motiverte. Det er ikke selvsagt hvordan anomalisering skal operasjonaliseres som inklusjonsgrunnlag i ordboksammenheng, men delstudie 1 og 3 indikerer til dels at en omhyggelig beregning av semantisk anomaliseringsgrad gir liten avkastning for mye arbeid. Delstudie 3 indikerer også at det kan være gunstig å inkludere sammensetninger på bakgrunn av anomali etter at man har kontrollert for diffusjon, slik at anomali reelt sett likevel er et nyttig inklusjonsgrunnlag blant sammensetninger med lav frekvens og ujevn spredning.

Anomalisering står dessuten i en motsetningsrelasjon til skjematisering (se forklaring av begge i underkapittel 2.7.1). I prinsippet vil en reint skjematisk sammensetning ha minimal anomaliseringsgrad, og vice versa. Felles for disse parametrene er at de fordrer et sammenlikningsgrunnlag. En sammensetning kan bare være skjematisk eller anomal i forhold til en viss mengde av andre sammensetninger. Dette åpner for at anomaliserings- og skjematiseringsgrad kan vurderes på ulike nivåer, altså med ulike mengder andre sammensetninger som sammenlikningsgrunnlag. For eksempel kan en sammensetning være anomal sammenlikna med norske sammensetninger generelt. *Kråketær* faller inn under denne klassifiseringa gitt momenta som trekkes fram i forrige avsnitt. Men en sammensetning kan også være anomal i forhold til for eksempel andre sammensetninger med samme for- eller etterledd, uten å være anomal sammenlikna med norske sammensetninger generelt. For eksempel skiller sammensetninga *gulvkald* seg fra de andre *gulv*-sammensetningene i BOB ved å være det eneste adjektivet. For en utvelgelsesmodell blir det da viktig å gjøre en avveining av hva en sammensetning eventuelt er anomal i forhold til. Terskelen bør her naturligvis være lavere for sammensetninger som stikker seg

ut i mengden av norske sammensetninger generelt, enn sammensetninger som stikker seg ut gitt en delmengde av sammensetninger med samme for- eller etterledd.

Skjematisering gir på sin side bare ordbokrelevans for sammensetninger som er skjematiske med hensyn til en mindre delmengde. Sammensetninger som er skjematiske i henhold til hele det sammensatte ordforrådet, får ingen økt ordbokaktualitet på bakgrunn av dette. I så fall ville dette kunne kvalifisere enhver sammensetning. Snarere kan det være aktuelt å innlemme sammensetninger som sammen viser et tydelig mønster innenfor en delmengde. I slike tilfeller vil brukeren kunne tilegne seg et produktivt skjema for fortolkning og produksjon av nye og liknende sammensetninger. Etter eksponering for sammensetningene *halsbetennelse* og *ørebetennelse* kan en språkbruker med letthet tolke og produsere analoge sammensetninger som *bihulebetennelse* og *lungebetennelse*. Enkelte sammensetninger kan såleis betegnes som usualiserte ved at de følger et etablert og produktivt skjema. Slike usualiserte skjema bør ideelt sett være representert i allmennordbøker, for eksempel ved at de mest diffunderte eksemplara innlemmes i ordboka. Mindre diffunderte eksemplarer kan dermed utelates. Såleis kan eksistensen av etablerte skjema innenfor delmengder av sammensetninger gi grunnlag for å innlemme visse sammensetninger og utelate visse andre.

Usualiseringsdomenet kan stundom motivere ordbokføring. Sammensetninger som er usualisert innenfor allment tilgjengelige domener som skole, politikk, natur, miljø, helse og liknende, kan være mer nærliggende ordbokkandidater enn sammensetninger innenfor mindre tilgjengelige domener som arkeologi, rallycross, hekling og undervannspolo. Videre er det et særlig ansvar for norske ordbøker å inkludere sammensetninger knytta til norsk historie, kultur, politikk og samfunnsliv. Selv om dette virker helt innlysende, fins det ingen soleklare grenseoppganger mellom relevante og irrelevante domener. I tvilstilfeller kan man gjøre en samla vurdering av domenets status i den (bredt forstått) norske virkeligheten, og den aktuelle sammensetningas bruksomfang innenfor domenet. Om domenet er relevant, kan man akseptere usualiserte, men mindre diffunderte sammensetninger, men om domenet har tvilsom relevans, trenger muligens sammensetninga også en anselig diffusjon for å kvalifisere til ordbokoppslag. I alle fall er det slik at usualiseringsdomene kan være et inklusjonsgrunnlag for sammensetninger som ikke kvalifiserer på annet grunnlag, gitt selvfølgelig at sammensetninga er tilstrekkelig usualisert.

Blant andre Zenner et al. (2014) skiller mellom kommunikativ og erfaringsbasert innprenting. Mentale begreper og deres betegnelser kan aktiveres jevnlig i minnet ved at man blir eksponert for eller bruker betegnelsen, men de kan også aktiveres ved eksponering for et ekstralingvistisk fenomen, jf. eksempelet i underkapittel 2.2.2 om å ri på en kamel vs. å høre ordet *kamel*. Siden det langt ifra er noe en-til-en-forhold mellom språklige frekvenser og med hvilken hyppighet vi støter på ulike fenomener (se dessuten underkapittel 2.6.5), er det nødvendigvis slik at noen fenomener må være over- og un-

derrepresentert i korpus jamført med hvor ofte folk i gjennomsnitt blir eksponert for dem. En skulle for eksempel tro at gjennomsnittlige mennesker var oftere eksponert for dørhåndtak enn skilsmisser; likevel er leksemet SKILSMISSE omtrent dobbelt så frekvent som leksemet DØRHÅNDTAK i LBK. Poenget her er at visse sammensetninger betegner begrep som kan bli aktivert jevnlig i minnet til språkbrukerne uten at den korresponderende betegnelsen forekommer hyppig i tekst eller tale. Begrep som etter alt å dømme er sterkt innprenta gjennom erfaring, og som har en usualisert betegnelse, kan med fordel inkluderes i ordbøker.

Oppmerksomhetsverdi (etter Svanlunds (2009) *uppmärksamhetsvärde*) kan også påvirke etableringsgraden. Svanlund (2009, 193) påpeker hvordan innovative og billedlige orddannelser kan være mer iøynefallende enn regelbundne og motiverte. Det samme gjelder orddannelser som tilhører kontroversielle domener. Såleis er oppmerksomhetsverdi knytta til både anomaliseringsgrad og usualiseringsdomene. For orddannelser i en relativt tidlig etableringsfase er det ikke uvanlig at ordet i kontekst ledsages av tekstlige markører som anførselstegn eller modifikatorer som «såkalt». Dette kan ifølge Svanlund både leses som et tegn på svak etablering, at ordet ikke er velkjent nok til å kunne benyttes umarkert, eller som forhøyet oppmerksomhetsverdi, hvilket i tur kan øke etableringa til det aktuelle ordet.[2] Ord som en avsender tenker at mottakeren kan komme til feste seg ved, er presumptivt viktigere å kommentere enn ord som passerer under mottakerens radar.

Dersom en sammensetning inngår i videre sammensatte eller avleda orddannelser, kan dette være et tegn på usualisering og til dels diffusjon, hvilket kan motivere ordbokoppføring. Når sammensetninger opptrer som ledd i mer komplekse orddannelser, oppfører de seg på lik linje med rotord. Derfor kan sammensetninger som inngår i mange nye orddannelser, være gode kandidater for inkludering.

I det følgende foreslås en mest mulig konkret operasjonalisering av de ovenstående inklusjonskriteria i en norsk kontekst.

5.3.1 Forslag til prosedyre for utvelgelse av sammensetninger

Med utgangspunkt i variablene ovenfor formulerer jeg her et forslag til en stegvis prosedyre for leksikografisk seleksjon av sammensetninger til en norsk allmennordbok. Prosedyren avhenger av språkressurser som gjør en i stand til å detektere og måle diffusjonen til aktuelle sammensetninger. Siden språkressursene for bokmål er vesentlig større og mer utbygde enn for nynorsk, vil prosedyren være tilrettelagt ressursene som fins for bokmål per 2023.

[2] Anførselstegn eller andre markeringer kan naturligvis også forekomme av andre årsaker, som at ordet brukes utenfor sitt vanlige domene.

1. Definer en avgrensa ordmengde å selektere sammensetninger fra, for eksempel ord med en gitt ordstamme som forledd eller som etterledd.

2. Gjør et trunkert søk i LBK og registrer antall forekomster og spredning per domene og år for sammensetninger som overstiger n antall forekomster. n avhenger av hvor mange belagte sammensetninger det fins innenfor ordmengden, og hva slags kapasitet det aktuelle leksikografiske prosjektet har. I mindre ordmengder kan det være aktuelt å vurdere samtlige belagte sammensetninger; i større ordmengder må man sette en hensiktsmessig grenseverdi for hvor mange som skal vurderes. Sammensetningene som skal vurderes innenfor ordmengden, får heretter benevnelsen *kandidater*. Sammensetninger som blir inkludert, fjernes fra den gjeldende kandidatlista.

3. Inkluder sammensetninger med frekvens- og/eller spredningsverdier over en viss terskel. Terskelverdien avhenger av ordbokas størrelse. For en ordbok på opp mot 100 000 ord (som BOB), kan et snitt på 0,5 «Deviation of Proportions» (DP) for år og domene (se Gries (2008); Paulsen (2022)) være en hensiktsmessig terskelverdi for spredning (gitt at frekvensen overstiger f.eks. terskelverdien 0.1 per million ord (pmo)) og 1 pmo for frekvens (gitt at spredningsverdien overstiger f.eks. terskelverdien 0,2). Terskelverdiene må bestemmes ut fra en pragmatisk og intuitiv avveining av hvor mange sammensetninger man kan eller bør ta med fra en ordmengde.

4. Suppler den resterende lista med kandidater (som ikke ble inkludert i punkt 3) med sammensetninger fra andre aktuelle kilder, for eksempel *Store norske leksikon* og andre ordbøker eller ordlister

5. Gjør et trunkert søk i Nasjonalbibliotekets n-gramsøk for henholdsvis bøker og aviser (heretter hhv. *NBavis og NBbok*). Sammenlikn lista over de x mest frekvente sammensetningene i disse korpusa med lista over gjenværende kandidater og legg til eventuelle nye kandidater. Verdien til x avhenger av hvor mange sammensetninger man finner belegg på innenfor ordmengden.

6. For alle foreliggende kandidater, registrer frekvensen i NBbok og NBavis. Siden NB-korpusa, til forskjell fra LBK, ikke er lemmatiserte, må man sørge for at alle bøyningsvarianter av kandidatleksemene inngår i frekvensberegninga.

7. Inkluder kandidater som innenfor ordmengden har høy frekvensrangering i NBbok og NBavis, og som ikke allerede har blitt inkludert.[3] Hva som utgjør en tilstrekkelig høy frekvensrangering avhenger av hvor mange sammensetninger med høy utbredelse det fins innenfor ordmengda.

[3]Ideelt sett skulle man også vurdert spredninga i NB-korpusa. Per november 2023 er dette svært tidkrevende å gjøre siden spredningsmål foreløpig ikke er implementert i grensesnittet til NB.

8. Inkluder resterende kandidater som har en eller flere av følgende egenskaper:

   (a) den er semantisk eller morfologisk uregelmessig.

   (b) den inngår i et mønster som indikerer en form for skjematisering som ikke blir tilstrekkelig representert med de allerede inkluderte sammensetningene innenfor ordmengden.

   (c) den er usualisert innenfor et domene som bør være godt representert i en allmennordbok

   (d) den har en langt større erfaringsbasert innprenting enn det som framgår av korpusutbredelsen.

   (e) den har høy oppmerksomhetsverdi, dvs. virker iøynefallende gitt sitt innhold.

   (f) den inngår i mange nye orddannelser.

Den ovenstående prosedyren gjør en stegvis seleksjon av sammensetninger innenfor en mindre ordmengde. Stega er arrangert slik at kandidatene med høyest diffusjon blir inkludert først, før andre kvalitative indikasjoner på etableringsgrad blir vurdert for de resterende kandidatene. Denne rekkefølgen bidrar til at leksikografen slipper å gjøre tidkrevende kvalitative vurderinger av samtlige kandidater, men heller kan bruke de kvalitative egenskapene til å håndplukke gode kandidater som ikke blir fanga opp i diffusjonsberegninga. Prosedyren har dessuten innebygde mekanismer for å regulere mengden inntak avhengig av størrelsen på den aktuelle ordboka.

5.3.2 Eksempel på anvendelse av utvelgelsesprosedyren

I det følgende beskriver jeg hvordan den ovenstående utvelgelsesprosedyren kan anvendes på en ordmengde.

1. Den avgrensa ordmengden er sammensetninger som begynner med *kjærlighet* i bokmål.

2. Siden LBK gir treff på godt over 300 forskjellige sammensetningsleksem med *kjærlighet* som forledd, setter jeg en grenseverdi på 5 forekomster og beregner frekvensen og spredninga til alle sammensetninger med 5 eller flere forekomster. I dette eksempelet utgjør det 62 kandidater.

3. Deretter inkluderer jeg 6 sammensetninger som har høyere LBK-frekvens enn 1 pmo (og minst 0,3 i DP), og ytterligere 8 sammensetninger som har høyere spredningsverdier enn 0,5 DP (og minst 0,1 i frekvens pmo) (se tabell 5.2).

4. Lista over resterende kandidater supplerer jeg med *kjærlighetsknop* og *kjærlighetseple* etter et trunkert søk på *kjærlighet** i Store norske leksikon.

5. Deretter sammenlikner jeg lista over de 10 mest frekvente *kjærlighet*-sammensetningene i NBavis og NBbok med kandidatlista. Ingen nye kandidater blir lagt til lista.

6. Deretter måler jeg frekvensen til de nåværende kandidatene i NBavis og NBbok.

7. Jeg inkluderer sammensetningene *kjærlighetsaffære* og *kjærlighetsdrama* som hver har over 25 000 forekomster sammenlagt fra NBavis og NBbok. Dette tilsvarer tilnærmet 0,2 pmo.

8. Avslutningsvis gjør jeg en vurdering av de kvalitative egenskapene (nevnt i punkt 8a–f i lista ovenfor) til de resterende kandidatene. *Kjærlighetseple* og *-knop* inkluderes på bakgrunn av anomalisering, og *kjærlighetsgudinne* og *-gud* inkluderes på bakgrunn av at de er usualiserte innenfor domenet TRO OG LIVSSYN, som kan sies å ha viktig kulturell betydning (se tabell 5.2).

I punkt 7 i lista ovenfor opererer jeg med ulike frekvensterskler for LBK og NB-korpusa. Dette fordi det er stor forskjell på frekvensen til *kjærlighet*-sammensetningene i de ulike korpusa. Mens 6 kandidater overstiger terskelen på 1 pmo (som tilsvarer 100 forekomster) i LBK, er det kun *kjærlighetshistorie* som overstiger 1 pmo i NB-korpusa (1 pmo tilsvarer her 135 000 forekomster totalt). 1 pmo-terskelen blir derfor for høy i NB-korpuset, og jeg opererer i stedet med 0,2 pmo. Disse frekvenstersklene er bestemt med bakgrunn i en pragmatisk og intuitiv avveining av hvor mange sammensetninger som dermed inkluderes. Det fins med andre ord ingen garanti for at de gir et optimalt resultat med henblikk på søkestatistikk.

Det at NB-korpusa ikke er lemmatiserte, gjør at det kan være vanskelig å fastslå frekvensen til leksem der en eller flere av bøyningsformene er homografiske med bøyningsformer innenfor andre leksem. For eksempel veit man ikke om formen *sag* tilhører verbleksemet *sage* eller substantivleksemet *sag*. For *kjærlighet*-sammensetningene har imidlertid

Tabell 5.2 lister de tjue inkluderte *kjærlighet*-sammensetningene og viser hvilke(t) inklusjonsgrunnlag hver enkelt sammensetning har i henhold til prosedyren ovenfor. Diffusjonskolonnen er tredelt mellom henholdsvis frekvens og spredning i LBK og frekvens i NB. Som man kan se, inkluderes mange sammensetninger på bakgrunn av flere punkter, mens noen, sånn som *kjærlighetsaffære* og *kjærlighetsknop*, kun kvalifiserer på ett punkt.

Lista i tabell 5.2 avviker noe fra den gjeldende lista over *kjærlighet*-sammensetninger i BOB (per 07.08.2023). Fire sammensetninger, *kjærlighetsdrama*, *-eple*, *-knop* og *-scene* fins ikke i BOB, mens ni BOB-førte sammensetninger, *kjærlighetsbud*, *-bånd*, *-drikk*, *-eventyr*, *-frukt*, *-fylt*, *-hat*, *-måltid* og *-vise*, ikke selekteres av prosedyren ovenfor. Blant

Ord	Diffusjon			Anomali	Skjema	Domene	Erf.innpr.	Opp.-verdi	Orddann.
	LBK		NB						
	Frek	Spred	Frek						
-affære	-	-	x	-	-	-	-	-	-
-barn	-	x	-	-	x	-	-	-	-
-brev	x	x	x	-	x	-	-	-	-
-dikt	-	x	x	-	x	-	-	-	-
-drama	-	-	x	-	x	-	-	-	-
-eple	-	-	-	x	-	x	-	-	-
-erklæring	x	x	x	-	-	-	-	-	-
-evne	-	x	-	-	-	-	-	-	-
-forhold	x	x	x	-	-	-	x	-	-
-full	-	x	-	x	-	-	-	-	-
-gud	-	-	-	-	-	x	-	x	-
-gudinne	-	-	-	-	-	x	-	x	-
-historie	x	x	x	-	x	-	-	-	-
-knop	-	-	-	x	-	-	-	-	-
-liv	x	x	x	x	-	-	-	-	-
-løs	-	x	-	x	-	-	-	-	-
-roman	-	x	x	-	x	-	-	-	-
-sang	-	x	x	-	x	-	-	-	-
-scene	-	x	-	-	x	-	-	-	-
-sorg	x	x	x	-	-	-	-	-	-

Tabell 5.2: Tabell over inkluderte sammensetninger på *kjærlighet-* og deres inklusjonsgrunnlag

disse sammensetningene er det kun *kjærlighetsbånd* som har en betydelig søkeinteresse, og denne burde ideelt sett blitt plukka opp av den stegvise prosedyren. For å øke presisjonen til prosedyren bør den i framtidig forskning spesifiseres ytterligere, testes og revideres.

I dette kapittelet har jeg med utgangspunkt i tidligere forskning og funna i denne avhandlinga trukket fram en gruppe variabler som later til å være hensiktsmessige for å identifisere sammensetninger som inngår i et gunstig sammensetningsutvalg for allmennordbøker. I tillegg har jeg presentert et forslag til operasjonalisering av de aktuelle variablene gjennom å samle dem i en konkret og detaljert prosedyre. I neste og siste kapittel drøfter jeg denne avhandlingas innfallsvinkel og resultater, før jeg foreslår spørsmål til framtidig forskning på sammensetninger innenfor en leksikografisk kontekst. Kappa avsluttes deretter med en kort oppsummering av avhandlinga som helhet.

Kapittel 6

Avsluttende diskusjon og oppsummering

I forrige kapittel har jeg samla trådene fra de teoretiske forutsetningene som presenteres i kapittel 2, tidligere forskning og funn og indikasjoner fra delstudiene for å beskrive et forslag til leksikologisk og leksikografisk rammeverk for utvelgelse og rangering av sammensetninger. Den stegvise prosedyren er beskrevet i detalj og eksemplifisert på et sammensetningsstrekk. I det følgende drøftes de mest sentrale avveiningene som denne avhandlinga og dens delstudier hviler på, og resultata de har avstedkommet.

6.1 Drøfting

Denne avhandlinga og dens delstudier har i større grad vektlagt identifisering, praktisk operasjonalisering og evaluering av etableringsvariabler enn litteraturgjennomgang og utvikling av nye begreper og kategorier. I stedet for å systematisk spore og kategorisere alle nyanser i det sammensatte ordtilfanget, har jeg gjort noen mer eller mindre tilfeldige nedslag i alfabetiske strekk for så å utvikle løsninger og gjøre målinger for de konkrete sammensetningseksemplara som da har kommet under lupen. Formålet med dette har vært å teste både validiteten og brukervennligheten til ulike hypotetiske utvalgsmetoder. Dette gjøres i både delstudie 1 og 3 og i forrige kapittel. Delstudie 2 skiller seg ut ved at den ikke inneholder noen rangering av sammensetninger, men snarere evaluerer reliabiliteten og validiteten til ulike korpusmetoder som mål på diffusjon. Det er essensielt for leksikografiske utvelgelsesmodeller at de har metoder for å gjøre treffsikre beregninger av diffusjon, hvilket delstudie 2 i stor grad bidrar til.

En fordel med den valgte innfallsvinkelen for denne avhandlinga er at den kulminerer i en gjennomtesta og empirisk belagt tilnærming til utvelgelse av sammensetninger. Disse punkta bidrar til at løsningene som presenteres, er prinsipielt og praktisk anvendelige for konkrete leksikografiske prosjekter. Med en mer teoritung innfallsvinkel kunne man endt opp med løsninger som kunne vært upåklagelige fra en lingvistisk synsvinkel, men

som ville fortona seg som lite gjennomførbare i møte med leksikografisk praksis.

En ulempe med den valgte innfallsvinkelen er at det ikke vinnes så mye ny innsikt om de relevante begrepa. For eksempel blir etableringsvariablene som presenteres i delkapittel 5.3, ikke uttømmende utforska og nyansert. Dermed etterlates spørsmål knytta til for eksempel hvordan man avgrenser anomalisering og usualiseringsdomene, hvordan man på en objektiv måte beregner erfaringsbasert innprenting og oppmerksomhetsverdi, og hvordan man identifiserer et skjema.

Etableringsvariablene som er presentert gjennom denne avhandlinga, har til dels blitt evaluert på bakgrunn av sin evne til å predikere søkeinteresse. Dette er på den ene siden en hensiktsmessig evalueringsvariabel siden sammensetninger som ingen søker på, typisk har mindre nytteverdi som oppslag enn de som blir søkt etter. På den andre siden er det ikke slik at leksikografer kan avstå fra å kartlegge deler av språkteigen bare fordi de ikke har noen klar evidens for at brukerne er interessert i alle deler av den. Dessuten foretas det mange søk i ordbøker på ord fra andre språk, som åpenbart ikke skal normeres bare fordi noen har søkt etter dem. Det er med andre ord ikke et mål for leksikografien at det skal være et en-til-en-forhold mellom lemmalista til ordboka og lemmalista til søkeloggen til ordboka. Snarere er det slik at søkeloggen kan validere de sammensetningsinntaka som er i tråd med søkeinteressen, mens andre sammensetningsinntak kan være vel så valide med utgangspunkt i at de inngår i den språkteigen ordboka søker å beskrive. Idealet for sammensetninger må likevel være at det fins en hensiktsmessig balanse mellom ubesøkte oppslagsord og lakuner, der sistnevnte er et større problem enn førstnevnte. Ordboka bør med andre ord forsøke å ha så høy sensitivitet som mulig – at den fanger opp reelt ettersøkte sammensetninger – selv om det skulle gå noe på bekostning av spesifisiteten, altså at antallet ubesøkte oppslagsord er relativt høyt.

Avhandlinga har gjennomgående fremma kombinasjonen av korpusfrekvens og korpusspredning som et valid og reliabelt mål på ususfrekvens. Korpusspredning kan operasjonaliseres på mange ulike måter (se oversikt i Gries (2008)) og har her blitt operasjonalisert via formlene *Deviation of Proportions* (*DP*) (fra Gries (2008)) og Juillands D (fra Juilland et al. (1970)). Begge disse formlene har sine svakheter. Gries (2021) kritiserer DP for å være for sterkt påvirket av korpusfrekvens. DP genererer verdier mellom 0 og 1 og måles over en predefinert oppdeling av korpuset, for eksempel basert på en metavariabel som årstall eller domene. Gries' argument er at DP-formelen som introduseres i Gries (2008), og som er brukt på ulike måter i denne avhandlinga, ikke kontrollerer for det faktum at det for særlig lavfrekvente n-gram er matematisk umulig å oppnå verdien 0 eller 1 i spredning, gitt antall korpusdeler. For å ta et banalt eksempel kan man se for seg at man måler spredninga til et hapax legomenon over to jevnstore korpusdeler. Både den pessimale og den optimale spredninga til dette n-grammet vil da nødvendigvis være 0,5. Spredningsverdien blir i dette tilfellet sterkt influert av det faktum at det bare

er én forekomst. For å bøte på dette problemet, og for å løsrive spredningsformelen fra korpusfrekvensen, foreslås det i Gries (2021) å justere DP-verdien til spesielt lavfrekvente n-gram slik at spredninga beregnes ut fra hva som er matematisk mulig gitt antall forekomster n-grammet har.

Juillands D har særlig den svakheten at pessimal spredning tildeles alle n-gram som kun forekommer i én korpusdel, uavhengig av hvor stor denne korpusdelen er. Selv om den aktuelle korpusdelen utgjør 95 % av korpuset, vil et n-gram som kun har forekomster i denne, få pessimal spredning.

Med andre ord må det tas et visst forbehold knytta til spredninga som i denne avhandlinga har blitt beregna over år og domene i LBK i delstudie 2 og 3, og over år i søkeloggen til standardordbøkene i delstudie 3. I delstudie 3 har imidlertid spredningsmålet vært konfigurert som snittet av DP og Juillands D over år og domene. I dette målet blir det noenlunde justert for svakheten som påpekes ovenfor, ved at de pessimale verdiene av Juillands D blir moderert av DP. Det er riktignok verdt å merke seg at svakhetene som hefter ved DP og Juillands D, i størst grad påvirker de mest lavfrekvente n-gramma. Disse får nemlig en noe lavere spredningsverdi enn deres distribusjon strengt tatt skulle tilsi. Det er imidlertid ingen grunn til å tro at resultata i avhandlinga er nevneverdig påvirket av disse svakhetene ved spredningsmålene. Siden de mest lavfrekvente n-gramma uansett ikke vil kvalifisere på bakgrunn av korpusutbredelse, er det trolig ingen ordbokkandidater som har blitt urettmessig ignorert av utvelgelsesprosedyrene i enten delstudie 3 eller i kappa på bakgrunn av svakheter i spredningsberegninga.

Utgangspunktet for denne avhandlinga var blant annet et ønske om å nyansere kriteria som blir anvendt for å selektere sammensetninger til ordbøker. Tradisjonelt har denne seleksjonen vært eksplisitt knytta til ususfrekvens og semantisk gjennomsiktighet (Fjeld & Vikør, 2008, 160), men om man studerer lemmalista i standardordbøkene, er det tydelig at det reelt sett har vært langt flere variabler i spill i utforminga av disse. På ett nivå står de tradisjonelle kriteria uforandra etter funna i denne avhandlinga. Det er det bruksmessig etablerte og det lingvistisk uetablerte som er forsøkt fanga av de anbefalte variablene her. Med andre ord anbefales det at ordbøkene etterlater en språklig utmark bestående av sammensetninger som er uetablerte i språkbruk på grunn av enten lav diffusjon eller usualisering innenfor et perifert domene, og som i tillegg følger helt ordinære morfosemantiske mønstre.

Den språklige innmarken er da presumptivt tilpassa både den produktive og reseptive bruken av ordboka. En bruker som konsulterer ordboka i forbindelse med skriving, kan like gjerne ha behov for informasjon om hvordan sammensetninga staves – eller om den i det hele tatt er en etablert ordform – som informasjon om hva sammensetninga betyr. I denne sammenhengen er det derfor relevant å ta med de sammensetningene som brukeren med størst sannsynlighet vil ha behov for å bruke når hen produserer norske

ytringer, uavhengig av hvor gjennomsiktige de er. I en reseptiv kontekst, hvor en bruker for eksempel kommer over et ukjent ord i en tekst, er det relevant å dekke inn de sammensetningene som en bruker med en viss sannsynlighet vil lure på betydninga til. I en reseptiv kontekst er det rimeligvis ikke like stort behov for å undersøke formelle egenskaper som staving og bøyning. Om samtlige ordbokbrukere var avanserte, voksne språkbrukere, kunne man kanskje tenke seg at det bare er de ugjennomsiktige sammensetningene som ettersøkes i reseptive kontekster, uavhengig av ususfrekvens. Realiteten for for eksempel standardordbøkene er imidlertid at de også benyttes hyppig av forskjellige innlærergrupper, deriblant skoleelever og andrespråksbrukere. Med dette i mente er det også viktig med tanke på den reseptive ordbokbruken å innlemme de sammensetningene som brukerne med størst sannsynlighet blir eksponert for.

6.2 Videre forskning

En nærliggende og potensielt løfterik videreutvikling av studiene som er gjort i denne avhandlinga, vil være videre analyser av søkestatistikk med flere prediktorer og mer nyanserte responsvariabler. Responsvariabelen oppslagsregularitet kan med fordel nyanseres slik at den inneholder informasjon som indikerer noe om hvorfor brukeren har søkt opp ordet. For eksempel går det an å hente ut informasjon om brukeren har trykka fram visninga av bøyningsparadigme eller ei. Som omtalt i forrige avsnitt er det fullt mulig å se for seg at ulike sammensetninger blir søkt opp av ulike grunner, og det kan derfor også tenkes at ulike variabler predikerer ulike typer søkeatferd. Det ville gitt mening om antatt gjennomsiktige sammensetninger først og fremst ble søkt etter av formmessige grunner, altså for å undersøke staving og bøyning, mens ugjennomsiktige sammensetninger oftere ble konsultert for å undersøke betydning. Denne hypotesen kan undersøkes med mer detaljert informasjon om brukerens søking i ordbøkene.

I tillegg hadde det vært gunstig med framtidige studier som undersøker flere prediktorer opp mot søkeatferd. For eksempel kunne datering for første belegg, ordlengde, formell anomalisering, spredning og frekvens i flere korpus, usualiseringsdomene, oppmerksomhetsverdi og videre orddanning blitt anvendt for å forsøke å forklare mer av variasjonen blant nominale sammensetninger. Man kunne også med fordel eksperimentert med ulike operasjonaliseringer av frekvens som for eksempel frekvens per ord i samme ordklasse, frekvens per frase, frekvens per dokument og liknende, eller brukt et regularitetsmål som indikerer både frekvens og spredning.

Et annet spor for den videre forskninga på leksikografi og sammensetninger er den psykolingvistiske forskninga. Her vil det være gunstig å undersøke sammenhengen mellom reaksjonstider og for eksempel anomaliseringsgrad, korpusspredning og usualiseringsdo-

mene.

Til slutt trengs det naturligvis videre forskning på kvalifiserende grunnlag for ordbokoppføring. Prosedyren som foreslås i forrige kapittel, behøver suksessiv foredling gjennom studier av den prediktive kapasiteten til dens variabler og deres operasjonaliseringer.

6.3 Oppsummering

Totalt sett kan det vitenskapelige bidraget til denne avhandlinga og arbeidet som ligger til grunn for den, fordeles på tre kategorier: 1) styrking av generell korpusmetodikk, 2) evidensbaserte innsikter om sammensetninger og søkeinteresse og 3) utvikling av prosedyrer for leksikografisk behandling av sammensatte ord. 1) og 2) er på mange måter essensielle fundamenter for 3), samtidig som de har overføringsverdi til henholdsvis korpuslingvistikk og leksikografisk forskning mer generelt.

Denne avhandlinga består av tre delstudier og en sammenfattende kappe. Delstudiene er alle utgitt som selvstendige artikler. Delstudie 1 gjør en utdyping av fenomenet semantisk anomalisering, delstudie 2 består av en kryssvalideringsanalyse av ulike korpusmåls evne til å gjenspeile utbredelse i usus, mens delstudie 3 undersøker korpusvariablene frekvens og spredning pluss en rekke lingvistiske variablers evne til å forklare variasjon i søkestatistikk for å utlede en modell for utvelgelse av sammensetninger. Bakgrunnen og funna fra disse delstudiene rammes inn, sammenfattes og diskuteres videre i denne kappa. Kappa og prosjektet for øvrig har tatt utgangspunkt i spørsmålet «Hva utgjør et gunstig sammensetningsutvalg i allmennordbøker, og hvilke metoder og variabler er hensiktsmessige for å identifisere medlemmer av dette utvalget?». Denne besvares delvis gjennom delstudiene og endelig gjennom et forslag til prosedyre her i kappa.

Referanser

Atkins, B.T Sue & Rundell, Michael. (2008). *The Oxford Guide to Practical Lexicography.* Oxford University Press.

Bakken, Kristin. (1998). *Leksikalisering av sammensetninger: en studie av leksikaliseringsprosessen belyst ved et gammelnorsk diplommateriale fra 1300-tallet* (Doktoravhandling). Universitetet i Oslo.

Bakken, Kristin. (2006). Lexicalization. I Keith Brown (red.), *Encyclopedia of language & linguistics* (s. 106–108). Elsevier Pergamon.

Bakken, Kristin & Vikør, Lars S. (2011). Samansette preposisjonar i norske dialektar. *Norsk Lingvistisk Tidsskrift*, *29*, 191–204.

Bauer, Laurie. (1983). *English Word-formation.* Cambridge.

Benczes, Réka. (2006). *Creative Compounding in English: The Semantics of Metaphorical and Metonymical Noun-noun Combinations* (vol. 19). John Benjamins Publishing.

Bergenholtz, Henning & Bøgelund, Christina. (2002). Hvor præskriptiv er en deskriptiv ordbog? Hvor deskriptiv er en præskriptiv ordbog? *LexicoNordica*(9), 79–106.

Borque, Yves Stephen. (2014). *Toward a typology of semantic transparency: The case of french compounds* (Doktoravhandling). University of Toronto, Canada.

Bundgaard, Peer, Ostergaard, Svend & Stjernfelt, Frederik. (2006). Waterproof fire stations? Conceptual schemata and cognitive operations involved in compound constructions. *Semiotica*, *2006*(161), 363–393. doi: 10.1515/sem.2006.071

Bäckerud, Erik, Nilsson, Pär & Sköldberg, Emma. (2020). Så används svenska akademiens ordböcker på nätet. Implicit och explicit feedback från användarna. *Nordiske Studier i Leksikografi*(15), 91–101. Hentet fra `https://tidsskrift.dk/nsil/article/view/124011`

Bøe, Maria Vetleseter. (2020). Frekvens. I Erik Bolstad (red.), *Store medisinske leksikon.* Hentet fra `https://snl.no/frekvens` (Lest 05.12.2023)

De Smedt, Koenraad. (2021). Smittsomme koronaord. *Oslo Studies in Language*, *11*(2), 59–73. doi: 10.5617/osla.8488

Downing, Pamela. (1977). On the creation and use of English compound nouns. *Language*, *53*(4), 810–842.

Egbert, Jesse & Burch, Brent. (2022). Which words matter most? Operationalizing

lexical prevalence for rank-ordered word lists. *Applied Linguistics*, 103–126. doi: 10.1093/applin/amac030

Eiesland, Eli Anne. (2015). *The semantics of Norwegian noun-noun compounds : a corpus-based study* (Doktoravhandling). University of Oslo, Oslo.

Eik, Ragnhild. (2019). *The morphosyntax of compounding in Norwegian* (Doktoravhandling). NTNU, Trondheim.

Ellis, Nick. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188. doi: 10.1017/s0272263102002024

Evans, Vyvyan & Green, Melanie. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh University Press.

Faarlund, Jan Terje, Lie, Svein & Vannebo, Kjell Ivar. (1997). *Norsk referansegrammatikk*. Universitetsforlaget.

Fauconnier, Gilles & Turner, Mark. (2008). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Fjeld, Ruth Vatvedt. (2002). Normering i klemme mellom språkteknologiske og pedagogiske ordbøker. *LexicoNordica*(9), 131–148.

Fjeld, Ruth Vatvedt, Nøklestad, Anders & Hagen, Kristin. (2020). Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk. I Janne Bondi Johannessen & Kristin Hagen (red.), *Leksikografi og korpus. En hyllest til Ruth Vatvedt Fjeld.* (vol. 11, s. 47–59). Oslo; Norway: Oslo Studies and Language.

Fjeld, Ruth Vatvedt & Vikør, Lars S. (2008). *Ord og ordbøker*. Høyskoleforlaget.

Fløgstad, Guro Nore. (2022). Språket som system vs. språket i bruk. *Norsk lingvistisk tidsskrift*, *40*(2), 187–212.

Frøyland, Rolf. (2012). Bjørn Eidsvåg kaster seg på kjendisvinbølgen. *Stavanger aftenblad*. Hentet fra `https://www.aftenbladet.no/kultur/i/Ge3bB/bjoern-eidsvaag-kaster-seg-paa-kjendisvinboelgen` (Lest 09.11.2023)

Gagné, Christina L & Shoben, Edward J. (1997). Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of experimental psychology: Learning, memory, and cognition*, *23*(1), 71–87.

Gardiner, Sir Alan. (1954). *The Theory of Proper Names*. London: Oxford University Press.

Geeraerts, Dirk. (2000). Salience phenomena in the lexicon. A typology. I Liliana Albertazzi (red.), *Meaning and Cognition: A Multidisiplinary Approach* (s. 79–102). John Benjamins Publishing Company. doi: 10.1075/celcr.2.05gee

Geeraerts, Dirk. (2006). *Words and Other Wonders: Papers on Lexical and Semantic Topics*. Mouton de Gruyter.

Gries, Stefan T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437. doi: 10.1075/ijcl.13.4.02gri

Gries, Stefan T. (2010). Dispersions and adjusted frequencies in corpora: further explo-

rations. I Stefan T. Gries, Stefanie Wulff & Mark Davies (red.), *Corpus Linguistic Applications: Current Studies, New Directions* (s. 197–212). Rodopi.

Gries, Stefan T. (2021). What do (most of) our dispersion measures measure (most)? dispersion? *Journal of Second Language Studies*, 171–205. doi: 10.1075/jsls.21029.gri

Gries, Stefan T. (2022a). On, or against?, (just) frequency. I Hans C. Boas (red.), *Directions for Pedagogical Construction Grammar* (vol. 49, s. 47–72). De Gruyter Mouton. doi: 10.1515/9783110746723-002

Gries, Stefan T. (2022b). Toward more careful corpus statistics: Uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics*, *1*(1). doi: 10.1016/j.rmal.2021.100002

Jarema, Gonia, Busson, Céline, Nikolova, Rossitza, Tsapkini, Kyrana & Libben, Gary. (1999). Processing compounds: A cross-linguistic study. *Brain and Language*, *68*(1), 362–369. doi: 10.1006/brln.1999.2088

Jarvad, Pia. (1995). Kryptosammensætninger og skabsafledninger i virkeligheden og i ordbøgerne. *Nordiske Studier i Leksikografi*(3), 221–231.

Jeanne d'arc – wikipedia. (2023). Hentet fra `no.wikipedia.org/wiki/Jeanne_d'Arc` (Lest 01.17.2023)

Johannessen, Janne Bondi. (2001). Sammensatte ord. *Norsk lingvistisk tidsskrift*, *1*, 59–92.

Johannessen, Janne Bondi. (2017). Rødøyd og bøyd: Samdanning og perfektum partisipp har parallell semantikk. *Norsk lingvistisk tidsskrift*, *35*(1), 27–47.

Juilland, Alphonse, Brodin, Dorothy R & Davidovitch, Catherine. (1970). *Frequency Dictionary of French Words*. Mouton de Gruyter.

Kinn, Torodd. (2014). Verbalt presens partisipp. *Norsk Lingvistisk Tidsskrift*, *32*(1). Hentet fra `https://ojs.novus.no/index.php/NLT/article/view/159`

Kjelsvik, Bjørghild. (2017). Samansetningar i norsk og leddeling i ordbøkene. I Oddrun Grønvik & Bjørghild Kjelsvik (red.), *Målreising og morsmålsdokumentasjon : heidersskrift til Oddrun Grønvik ved 70-årsleitet* (s. 145–171). Oslo: Novus.

Kulbrandstad, Lars Anders & Kinn, Torodd. (2016). *Språkets mønstre*. Universitetsforlaget.

Kåss, Erik. (2020). Frekvens. I Erlend Hem (red.), *Store medisinske leksikon.* Hentet fra `https://sml.snl.no/frekvens` (Lest 05.12.2023)

Lakoff, George & Johnson, Mark. (1984). *Metaphors We Live By.* University of Chicago Press.

Langacker, Ronald W. (1987). *Foundations of Cognitive Grammar. Vol. 1: Theoretical Prerequisites*. Stanford University Press.

Laufer, Batia & Nation, Paul. (1995). Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, *16*(3), 307-322. doi: 10.1093/applin/

16.3.307

Leech, Geoffrey. (2007). New resources, or just better old ones? The Holy Grail of representativeness. I Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (red.), *Corpus Linguistics and the Web* (vol. 59, s. 133–149). Leiden, The Netherlands: Brill. doi: 10.1163/9789401203791_009

Leira, Vigleik. (1992). *Ordlaging og ordelement i norsk*. Det Norske Samlaget.

Leksikografisk bokmålskorpus. (2023). Distributed by the CLARINO. UiB Portal: hdl:11495/E1A4-54BE-FCD4-1.

Levshina, Natalia. (2015). *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. John Benjamins Publishing Company.

Levshina, Natalia. (2020). Conditional inference trees and random forests. I Stefan T Gries & Magali Paquot (red.), *A Practical Handbook of Corpus Linguistics* (s. 611–643). Springer International Publishing. doi: 10.1007/978-3-030-46216-1_25

Libben, Gary. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, *61*(1), 30–44. doi: 10.1006/brln.1997.1876

Libben, Gary, Gibson, Martha, Yoon, Yeo Bom & Sandra, Dominiek. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, *84*(1), 50–64.

Lieber, Rochelle & Štekauer, Pavol. (2009). *The Oxford Handbook of Compounding*. Oxford University Press.

Loenheim, Lisa. (2019). *Att tolka det sammansatta. Befästning och mönster i första-och andraspråkstalares tolkning av sammansättningar* (Doktoravhandling). Göteborgs Universitet.

Maguire, Phil, Devereux, Barry, Costello, Fintan & Cater, Arthur. (2007). A reanalysis of the CARIN theory of conceptual combination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 811–821. doi: 10.1037/0278-7393.33.4.811

Maguire, Phil, Maguire, Rebecca & Cater, Arthur W.S. (2010). The influence of interactional semantic patterns on the interpretation of noun–noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 288–297. doi: 10.1037/a0018687

McDonald, Scott & Shillcock, Richard. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and speech*, *44*(Pt 3), 295–323. doi: 10.1177/00238309010440030101

McIntyre, Michael. (2019). *Simplifying English for the Americans | Michael Mcintyre - Youtube.* Hentet fra `https://www.youtube.com/watch?v=UCo0hSFAWOc` (Sett 01.16.2023)

Mehl, Seth. (2016). *Corpus onomasiology: A study in World Englishes* (Doktoravhand-

ling). University College London.

Mellenius, Ingmarie. (1997). *The acquisition of nominal compounding in Swedish* (Doktoravhandling). Lund University.

Merriam-Webster.com Dictionary, Merriam-Webster. (2023). *«blackbird».* Hentet fra `https://www.merriam-webster.com/dictionary/blackbird` (Lest 05.12.2023)

Müller-Spitzer, Carolin, Wolfer, Sascha & Koplenig, Alexander. (2015). Observing online dictionary users: Studies using Wiktionary log files. *International Journal of Lexicography*, *28*(1), 1–26. doi: 10.1093/ijl/ecu029

NAOB = Det Norske Akademis ordbok. (2023). Oslo: Det Norske Akademi for Språk og Litteratur. Hentet fra `https://naob.no` (Lest 05.12.2023)

Nasjonalbiblioteket. (2023). *Nasjonalbiblioteket / n-gram.* Hentet fra `https://www.nb.no/ngram`

Nesset, Tore. (2011). Metafor og metonymi: personkarakteriserende sammensatte substantiv i norsk. *Maal og Minne*, *101*(1), 32–64.

Nesset, Tore. (2017). Spøkelsesfiske, makrellfotball og traktoregg: norske sammensetninger og konseptuell integrasjon. *Maal og Minne*, *108*(2), 85–110.

Nesset, Tore. (2018). Metakonstruksjonssammensetninger. *Maal og Minne*, *110*(1), 139–164.

Nesset, Tore & Sokolova, Svetlana. (2019). Compounds and culture: Conceptual blending in Norwegian and Russian. *Review of Cognitive Linguistics*, *17*(1), 257–274. doi: 10.1075/rcl.00034.nes

Norsk ordbank. (2023). Universitetet i Bergen. Hentet fra `www.uib.no/ub/spesialsamlingene/160646/ordb%C3%B8ker` (Lest 27.11.2023)

Ordbøkene / Bokmålsordboka og Nynorskordboka. (2023). Språkrådet og Universitetet i Bergen. Hentet fra `https://ordbokene.no`

Paulsen, Mikkel Ekeland. (2022). Assessing word commonness: Adding dispersion to frequency. *International Journal of Corpus Linguistics*, *28*, 318–343. doi: 10.1075/ijcl.21037.eke

Paulsen, Mikkel Ekeland. (2023). Wheat or chaff? A compound selection model based on look-up data. *International Journal of Lexicography*, 306–324. doi: 10.1093/ijl/ecad013

Pepper, Steve. (2020). *The typology and semantics of binominal lexemes* (Doktoravhandling, Universitetet i Oslo). Hentet fra `https://www.hf.uio.no/iln/forskning/aktuelt/arrangementer/disputaser/2020/pepper_avhandling_trykkversjon.pdf`

Piantadosi, Steven T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychon Bull Rev*, *21*, 1112–1130. doi: 10.3758/s13423-014-0585-6

Pilke, Nina. (2008). Ordboksanvändning hos blivande språkexperter och hos professio-

nella översättare. *LexicoNordica*(15), 135–154.

Pripp, Are. (2020). Kryssvalidering – Å analysere dataene på kryss og tvers. *Tidsskrift for Den norske legeforening*. doi: 10.4045/tidsskr.20.0154

Revisjonsprosjektet. (2023). Hentet fra `https://www.uib.no/lle/revisjonsprosjektet` (Lest 09.11.2023)

Rummelhoff, Eirik-Mathias Bjørnø & Frøslie, Kathrine Frey. (2023). Frekvens (statistikk). I Erik Bolstad (red.), *Store norske leksikon*. Hentet fra `https://snl.no/frekvens_-_statistikk` (Lest 05.12.2023)

Ryder, Mary. (1989). *Ordered Chaos: A Cognitive Model for the Interpretation of English Noun-noun Compounds*. ProQuest Dissertations Publishing. Hentet fra `http://search.proquest.com/docview/303670949/`

Sakshaug, Laila. (1999). *Norwegian compound deverbal nouns : An autolexical analysis in morphology, syntax, and semantics* (Doktoravhandling). NTNU, Trondheim.

Sandra, Dominiek. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology Section A*, *42*(3), 529–567. doi: 10.1080/14640749008401236

Savický, Petr & Hlavácová, Jaroslava. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*(3), 215–231. doi: 10.1076/jqul.9.3.215.14124

Scalise, Sergio & Vogel, Irene. (2010). *Cross-disciplinary Issues in Compounding*. John Benjamins Publishing Company.

Schmid, Hans-Jörg. (2015). A blueprint of the entrenchment-and- conventionalization model. *Yearbook of the German Cognitive Linguistics Association*, *3*(1), 3–26. doi: 10.1515/gcla-2015-0002

Schmid, Hans-Jörg. (2020). *The Dynamics of the Linguistic System. Usage, Conventionalization, and Entrenchment*. Oxford University Press. doi: 10.1093/oso/9780198814771.001.0001

de Schryver, Gilles-Maurice, Joffe, David, Joffe, Pitta & Hillewaert, Sarah. (2006). Do dictionary users really look up frequent words? On the overestimation of the value of corpus-based lexicography. *Lexikos*, *16*, 67–83. doi: 10.4314/lex.v16i1.51504

Schäfer, Martin. (2018). *The Semantic Transparency of English Compound Nouns*. Language Science Press.

Shoben, Edward J & Gagné, Christina L. (1997). Thematic relations and the creation of combined concepts. I T.B Ward, S.M Smith & Vaid J (red.), *Creative Thought: An Investigation of Conceptual Structures and Processes* (s. 31–50). Washington, DC, US: American Psychological Association. doi: 10.1037/10227-002

Språkrådet. (2023). *Protokoll frå styremøte 7.juni 2023.* Hentet fra `www.sprakradet.no/globalassets/vi-og-vart/styreprotokoller/2023/protokoll-fra-styremote-7.-juni-2023.pdf` (Lest 29.10.2023)

Stefanowitsch, Anatol. (2020). *Corpus Linguistics. A Guide to the Metholodogy*. Berlin: Language Science Press.

Svanlund, Jan. (2002). Lexikalisering. *Språk och stil*, *12*, 7–45.

Svanlund, Jan. (2009). *Lexikal etablering: En korpusundersökning av hur nya sammansättningar konventionaliseras och får sin betydelse*. Acta Universitatis Stockholmiensis.

Štekauer, Pavol, Valera, Salvador & Körtvélyessy, Lívia. (2012). *Word-formation in the World's Languages: A Typological Survey*. Cambridge University Press.

Tagliamonte, Sali A. & Baayen, R.Harald. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135–178. doi: 10.1017/S0954394512000129

Taylor, John R. (2012). *The Mental Corpus. How language is represented in the mind*. New York: Oxford University Press.

ten Hacken, Pius. (2016). *The Semantics of Compounding*. Cambridge University Press. doi: 10.1017/cbo9781316163122

Theil, Rolf. (2016). Den prototypiske norske samansetninga. I Hans-Olav Enger, Monica I.Norvik Knoph, Kristian E Kristoffersen & Marianne Lind (red.), *Helt fabelaktig! Festskrift til Hanne Gram Simonsen på 70-årsdagen* (s. 235–253). Novus forlag.

Tomasello, Michael. (2006). Usage-based linguistics. I Dirk Geeraerts (red.), *Cognitive Linguistics: Basic Readings* (s. 439–459). De Gruyter, Inc.

Trap-Jensen, Lars, Lorentzen, Henrik & Sørensen, Nicolai H. (2014). An odd couple – Corpus frequency and look-up frequency: what relationship? *Slovenščina 2.0*, *2*(2), 94–113.

Wallis, Sean & Mehl, Seth. (2022). Comparing baselines for corpus analysis: Research into the get-passive in speech and writing. I Ole Schützler & Julia Schlüter (red.), *Data and Methods in Corpus Linguistics: Comparative Approaches* (s. 101–126). Cambridge University Press. doi: 10.1017/9781108589314.005

Winter, Bodo. (2020). *Statistics for Linguists - An Introduction Using R*. New York: Routledge.

Wolfer, Sascha, Koplenig, Alexander, Meyer, Peter & Müller-Spitzer, Carolin. (2014). Dictionary users do look up frequent and socially relevant words. Two log file analyses. I Andrea Abel, Chiara Vettori & Natascia Ralli (red.), *Proceedings of the 16th EURALEX International Congress* (s. 281–290). Bolzano, Italy: EURAC research.

Zenner, Eline, Speelman, Dirk & Geeraerts, Dirk. (2014). Core vocabulary, borrowability and entrenchment: A usage-based onomasiological approach. *Diachronica*, *31*(1), 74–105. doi: 10.1075/dia.31.1.03zen

Zimmer, Karl. (1971). Some general observations about nominal compounds. *Working Papers on Linguistic Universals, Stanford University*, *5*, C1–23. doi: 10.3987/

r-1987-05-1177

Zimmer, Karl. (1972). Appropriateness conditions for nominal compounds. *Working Papers on Linguistic Universals, Stanford University*, *8*, 3–20.

Vedlegg A

Delstudier

I Delstudie 1: Svartsjuk tankelesing på vandresafari — en modell for bedømmelse av sammensatte ords gjennomsiktighet

Svartsjuk tankelesing på *vandresafari* – en modell for bedømmelse av sammensatte ords gjennomsiktighet

Mikkel Ekeland Paulsen

Compounds are ubiquitous in Norwegian language use, and therefore also in Norwegian dictionaries. Whether a compound is selected for entry in the dictionary depends in part on the lexicographers' ability to distinguish transparent from nontransparent compounds. However, what decides if a compound is semantically transparent has never been fully investigated scientifically, at least not from a lexicographic point of view. This study outlines a model for assessing the semantic transparency of Norwegian compounds and presents the results of the model applied to a set of compounds.

1. Innledning

Sammensetninger utgjør en betydelig del av ordtilfanget i *Bokmålsordboka* (BOB) og *Nynorskordboka* (NOB), ifølge beregninger så mye som 60 % (Kjelsvik 2018). Et trunkert søk på ordet *arbeid* (arbeid*) i Leksikografisk bokmålskorpus (LBK; se også Knudsen & Fjeld 2013) gir treff på over 2700 ulike lemma. Noen av disse er usammensatte, enkelte er ren støy, mens over 90 % er sammensetninger. I moderne leksikografi utgjør tekstkorpus et av de viktigste kildematerialene. Her er utvelgelse av sammensatte oppslagsord notorisk problematisk. I norsk er orddanning via sammensetting høyproduktivt, og det er derfor viktig at leksikografer har velfunderte metoder for å vurdere sammensetningenes leksikografiske relevans.

Et viktig kriterium for å utelukke sammensetninger er semantisk gjennomsiktighet. Likevel fins det foreløpig ikke en velfundert

leksikografisk framgangsmåte for å bedømme slik gjennomsiktighet.[1]

I denne artikkelen presenterer og anvender jeg en modell for bedømmelse av gjennomsiktighet i sammensetninger, basert på faktorer som av ulike sammensetningsforskere hevdes å ha innflytelse på dette. Modellen anvendes til en undersøkelse av et utvalg sammensetninger som blir plassert på en skala fra minst til mest gjennomsiktig. Denne framgangsmåten tar sikte på å være et ledd i å utvikle en metodikk for å identifisere ordbokaktuelle sammensetninger.

2. Teoretisk bakgrunn

I det følgende gjør jeg rede for termen *gjennomsiktige sammensetninger* og påpeker problematiske sider ved den tradisjonelle definisjonen av denne. Deretter diskuterer jeg ulike egenskaper som hevdes å påvirke gjennomsiktighetsgraden til den enkelte sammensetning.

Fjeld & Vikør (2008) trekker opp et skille mellom såkalte gjennomsiktige sammensetninger, til dømes *teaterbillett*, *postfunksjonær* og *jarnomn*, og sammensetninger med spesialiserte betydninger som ikke framgår direkte av sammensetningsledda, til dømes *teatersport*, *postgiro* og *jarnbane*.[2] Distinksjonen bidrar til å skille ordbokaktuelle sammensetninger fra sammensetninger som kan utelates.

Gjennomsiktige sammensetninger og motstykket *ugjennomsiktige* (*opake*) sammensetninger er mye brukte termer innenfor sammensetningslitteraturen, og kanskje særlig innenfor praktisk

1 Det fins andre kriterier også, til dømes frekvens i bruk. Dessuten avhenger det av ordbokas formål hvor viktig kriteriet om semantisk gjennomsiktighet er.

2 Døma til Fjeld & Vikør er nynorskformer. *Jarnomn* og *jarnbane* tilsvarer *jernovn* og *jernbane* på bokmål.

leksikografi (se bl.a. Fjeld & Vikør 2008). Termparet begrepsliggjør sammensetninger som et objekt med sider som vi i varierende grad er i stand til å se igjennom (Svanlund 2002). De mest regelbundne og forutsigbare sammensetningene har angivelig mer gjennomskuelige flater enn de mindre regelbundne. For å avgjøre om en sammensetning er gjennomsiktig, bruker man ofte en addisjonsmetafor (se til dømes Bakken 1998 og Fjeld & Vikør 2008). Her framstilles betydninga til en sammensetning som en sum av ledda. Presumptivt blir da formelen *forledd* + *etterledd* = betydning: *jern* + *ovn* = *jernovn*. Motstykket er sammensetninger hvor denne formelen tilsynelatende ikke passer, til dømes *jern* + *bane* ≠ *jernbane*.

Svanlund (2002) kritiserer dette synet for å utgå fra et forenklet syn på ordsemantikk og å overse sammensetningenes betydningspotensial. *Jernovn* kan like gjerne bli brukt med betydninga 'ovn som man smelter jern i' (dette fins til dømes i bøker om jernproduksjon) som 'ovn av jern'. De er begge potensielle og aktuelle betydninger av sammensetninga, men dette kommer ikke fram når man behandler betydninga som om det var et enkelt regnestykke.

Når man vurderer om sammensetninger er gjennomsiktige, er det viktig å presisere at det man vurderer, er gjennomsiktighet på tvers av ulike språkbrukere og språklige kontekster. Det vil alltid være individuell variasjon i fortolkninga av sammensetninger (se til dømes Loenheim 2019 og Ryder 1994). Sammensetninger er prinsipielt mangetydige (Svanlund 2002) i den forstand at de via polysemi i forledd og etterledd og i relasjonen dem imellom kan anta et multiplum av gangbare betydninger. Likevel tyder mye på at noen sammensetninger *oppleves* som mer gjennomsiktige enn andre (Svanlund 2002 og Loenheim 2019). Det er til dømes grunn til å tro at en sammensetning som *adamseple* oppleves som mindre gjennomsiktig enn *treski*. Med dette menes det at det er vanskeligere å utlede den konvensjonelle betydninga til sammensetninga ut fra ledda i *adamseple* enn i *treski*. Hva denne forskjel-

len i opplevd gjennomsiktighet består i, kan blant annet forklares gjennom de faktorene jeg trekker fram her.

Svanlund (2002, 2009) presiserer forskjellen mellom *bedømmelse* og *opplevelse* av gjennomsiktighet. Opplevelse av gjennomsiktighet gjelder språkbrukerens første møter med en til da ukjent sammensetning. Leksikografer må som regel nøye seg med å bedømme gjennomsiktighet. Bedømmelsen skjer da utenfor kontekst og derfor reelt sett på tvers av ulike kontekster. Svanlund (2002) knytter kontekstfri gjennomsiktighet til *motiveringsgrad*. Motivering forstås her som en identifiserbar sammenheng mellom et ords betydning innenfor og utenfor en sammensetning. Man kan til dømes konstatere at betydninga av *blå* er å gjenfinne i sammensetninga *blåbær*, som tross alt denoterer et blått bær. Her framstår det ikke arbitrært, men snarere motivert at akkurat adjektivet *blå* og substantivet *bær* settes sammen for å denotere nettopp 'blått bær'. Såleis har sammensetninga *blåbær* høy motiveringsgrad. På den annen side burde *blåbær* med full gjennomsiktighet kunne bli brukt om alle blå bær, slik som den frie frasen *blått bær*. Den konvensjonelle betydninga av *blåbær* er imidlertid snevrere enn det som framgår av ledda alene, altså fins det en distinksjon mellom frasen *blått bær* og *blåbær* som gjør sistnevnte mindre gjennomsiktig, og derfor til en aktuell kandidat for ordbøker. Det vil si at en sammensetning kan ha høy motiveringsgrad, altså at leddas konvensjonelle betydninger er gjenfinnbare i sammensetningas konvensjonelle helhetsbetydning, men ikke være helt gjennomsiktig likevel. Førstnevnte er med andre ord en nødvendig, men ikke en tilstrekkelig betingelse for at en sammensetning skal bedømmes som helt gjennomsiktig.

Det at *blåbær* har en snevrere konvensjonell betydning enn ledda skulle tilsi, betegner Svanlund (2002) som at sammensetninga har tilleggselementer. Et annet døme er sammensetninga *kvinnesak*, der det på ingen måte er tilstrekkelig å vite betydninga av ledda for å komme fram til helheten. En stor gruppe sammen-

setninger som har tilleggselementer, er fagtermer. For disse er den konvensjonelle betydninga snevrere enn det som framgår av ledda alene, tenk bare på termen *sammensetning* i grammatikken.

En forutsetning for ideen om motiveringsgrad og tilleggselementer er at sammensetninger har tydelige konvensjonelle betydninger som språkbrukere i hovedsak er enige om. Dette er ikke en uproblematisk forutsetning. Schmid (2015) understreker at det er et samspill mellom konvensjonalisering og eksponeringshyppighet (eller innprenting) som leder fram til en gitt språkbrukers tolkning av et gitt uttrykk. I den grad det fins slike konvensjonelle betydninger, fins det ingen garanti for at leksikografer har tid eller ressurser til å avdekke disse i alle tilfeller. Spørsmålet om hvordan man fullt ut kan identifisere konvensjonelle betydninger, må imidlertid besvares annetsteds. I kapittel 3 beskriver jeg nærmere hvordan jeg identifiserer den konvensjonelle betydninga til uttrykka som inngår i undersøkelsen.

Et annet poeng fra Svanlund (2002) er at sammensetninger består av ledd som utgjør en minimal kontekst for hverandre. En sammensetning kan såleis bidra til å disambiguere seg selv ved at ledda spesifiserer hverandre. Et døme på dette har vi i *blåfarge*, der etterleddet *-farge* bidrar til å spesifisere at betydninga til forleddet *blå-* ikke er 'politisk konservativ' eller 'melankolsk', men snarere 'fargen blå'.

Shoben & Gagné (1997) har funnet at ulike ledd har *relasjonspreferanser* som språkbrukerne anvender når de tolker og produserer sammensetninger. Relasjonspreferansene stammer fra eksponeringshyppighet, dvs. hvor ofte vi produserer og blir eksponert for en bestemt relasjon, men også fra ekstralingvistisk kunnskap om strukturen til forskjellige substanser og prosesser (til dømes at jern er et materiale mange gjenstander består av). Relasjonspreferansene er altså en type forhåndskunnskap vi aktiverer i møte med nye sammensetninger. Til dømes bruker *fjell* som forledd å angi plasseringa til det som etterleddet betegner, som i *fjellgeit*. Konklusjo-

nen til Shoben & Gagné (1997) er at språkbrukerne har lettest for å tolke sammensetninger der relasjonen mellom for- og etterledd rimer med relasjonspreferansene til forleddet. Som omtalt i Loenheim (2019) har denne konklusjonen blitt utdypet av blant andre Maguire, Maguire & Cater (2010) og Svanlund (2009). De har vist at også etterleddets relasjonspreferanser spiller inn.

Loenheim (2019) finner at språkbrukere i hovedsak er mer samstemte i fortolkninga av frekvente sammensetninger, som de sannsynligvis har blitt eksponert for før, enn i fortolkninga av infrekvente sammensetninger. Når det er samstemmighet om forståelsen av infrekvente sammensetninger, skyldes det ofte at de tolkes i analogi med mer innarbeida sammensetninger med samme relasjonsmønster. *Politikvinne* kan til dømes tolkes i analogi med *politimann*. Gjennomsiktighet blir da også et spørsmål om hvorvidt det fins analoge sammensetninger som avkoderen har kjennskap til og kan generalisere ut fra.

Hittil har jeg diskutert gjennomsiktighet i sammensetninger uten å differensiere mellom ulike typer.[3] Substantiv-substantiv-sammensetninger er den vanligste typen i norsk, og den som med god margin er mest omtalt i litteraturen. Sammensetting begrenser seg likevel på ingen måte til denne typen. De leksikalske ordklassene – dvs. substantiv, verb, adjektiv, adverb og til en viss grad preposisjoner – kan med noen unntak kombineres fritt med hverandre (noen dømer er *balanseføre*, *nordetter*, *langveis* og *bofast*). Dette er en kjensgjerning for leksikografer, som ikke kan vie spesiell oppmerksomhet til én type sammensetninger. Selv om sammensetningstypen nok kan ha innvirkning på gjennomsiktigheten (Loenheim 2019), utelater jeg dette perspektivet her. Jeg kompenserer med å teste modellen på tre ulike typer forledd, som uunngåelig dekker minst tre ulike typer sammensetninger.

3 *Type* dreier seg her og videre i teksten om ordklassetilhørigheten til ledda. Sammensetninger kan også kategoriseres etter andre kriterier.

I opplevelsen av gjennomsiktighet gjør også distribusjonsaspekter seg gjeldende (Svanlund 2009 og Loenheim 2019). Disse aspektene faller imidlertid utenfor rammene av denne artikkelen og må heller sees i sammenheng med vurderinger av sammensetningers utbredelse i korpus.

3. Modell for bedømmelse av gjennomsiktighet

I dette kapitlet presenterer jeg en gruppe gjennomsiktighetsfaktorer som ble diskutert i kapittel 2, og beskriver en metode for hvordan disse faktorene kan brukes til å bedømme gjennomsiktighet i sammensetninger. Metoden tester jeg i kapittel 4 på en samling sammensetninger som jeg introduserer i slutten av dette kapitlet.

For å vurdere graden av gjennomsiktighet vekter modellen sammensetninga etter følgende kriterier:

- **tilleggselement**: hvorvidt fortolkninga av en sammensetning fordrer kunnskap om mer enn betydninga til ledda
- **motiveringsgrad**: hvorvidt ledda bidrar med sine konvensjonelle betydninger
- **disambiguering**: hvorvidt ledda disambiguerer hverandre
- **relasjonsmønster**: hvorvidt sammensetninga kan forstås ut fra leddas relasjonspreferanser
- **analogi**: hvorvidt sammensetninga kan forstås i analogi med mer frekvente sammensetninger

I det følgende forklarer jeg hver faktor nærmere.

3.1. Metode

Tilleggselement: Har sammensetninga konvensjonell(e) betydning(er) som er videre eller snevrere enn det som framgår av led-

da alene? **Framgangsmåte**: Sammensetning X får +1 hvis den på tilfredsstillende vis kan parafraseres uten at man tar i bruk andre innholdsord enn de som inngår i sammensetninga (sånn som *hårfjerning → fjerning av hår*).

Motiveringsgrad: Bidrar ledda med sine konvensjonelle betydninger i sammensetninga? Til dømes er en som er svartsjuk, verken konvensjonelt sjuk eller konvensjonelt svart. **Framgangsmåte**: For å operasjonalisere vurderinga av hva som er et gitt ledds konvensjonelle betydning, anser jeg som en nødløsning betydning 1 i BOB som den mest konvensjonelle. Om ordet ikke står i BOB, vil jeg konsultere *Det Norske Akademis ordbok* (NAOB).[4] Framgangsmåten blir slik: Sammensetning X får +1 hvis betydning 1 av forleddet er aktiv og + 2 hvis betydning 1 av etterleddet er aktiv i sammensetninga. Ellers får den 0. Kopulative sammensetninger, som ikke har en distinkt semantisk kjerne, kan kun få +1 for hvert ledd.

Disambiguering: Disambiguerer ledda hverandre? **Framgangsmåte**: Sammensetning X gis +1 i gjennomsiktighet hvis ledd A har flere betydningsnivåer i BOB og ledd B bidrar til å identifisere hvilket av disse betydningsnivåene av ledd A som er aktivt i sammensetninga, eller vice versa.

Relasjonsmønster: Er sammenstillinga av ledd A og B i tråd med ledd As eller ledd Bs relasjonspreferanser? Faktoren *relasjonsmønster* er inkludert i modellen med bakgrunn i psykolingvistiske studier av Shoben & Gagné (1997) og Maguire, Maguire & Cater (2010), som indikerer at relasjonspreferansene til ledda letter og stabiliserer tolkninga av sammensetningenes betydning. For å bruke et relasjonsrammeverk som er empirisk utviklet med basis i norsk, benytter jeg meg av Eieslands (2015) 14 relasjonstyper for

4 Det er ikke ideelt å bruke betydning 1 i NAOB, siden denne ordboka eksplisitt oppgir at betydning 1 er den *eldste* og ikke den mest konvensjonelle betydninga. Disse er stundom, men slettes ikke alltid, sammenfallende. Når NAOB likevel brukes i tillegg til BOB, er det i mangel av andre dugelige bokmålsressurser.

substantiv-substantiv-sammensetninger. Andre sammensetningstyper, som adjektiv-adjektiv, adjektiv-substantiv, substantiv-verb og adjektiv-verb, er viet langt mindre oppmerksomhet i litteraturen. Her er det derfor ikke mulig å ta i bruk et veldefinert rammeverk på samme måte. I sammensetninger der et verb er sammenstilt med et substantiv, dreier relasjonsmønsteret seg om den semantiske rollen til substantivet. Det samme gjelder når et verbavledet ledd er sammenstilt med et substantiv. *Folk* er til dømes agens i sammensetninga *folkelesing*, mens *avis* er patiens i *avislesing*. Når det gjelder adjektivsammensetninger, skiller Loenheim (2019) mellom kopulative og determinative adjektiv-adjektiv-sammensetninger. *Gulgrønn* kan til dømes tolkes kopulativt som 'gul og grønn' eller determinativt som 'gulaktig grønn'. I substantiv-adjektiv-sammensetninger skiller hun mellom en sammenliknende og en årsaksbeskrivende relasjon mellom ledda. *Iskald* ('kald som is') er døme på det første og *solvarm* ('varm av sola') på det andre. **Framgangsmåte**: Sammensetning X gis +2 i gjennomsiktighet hvis ledd A og ledd B har en relasjon til hverandre som er i tråd med begge leddas relasjonspreferanser. X gis +1 hvis ledd A eller ledd B har en relasjon til motsatt ledd som følger relasjonspreferansene. Preferansene utledes fra de fem mest frekvente sammensetningene i LBK med ledd A som forledd, og tilsvarende fra de fem mest frekvente med ledd B som etterledd. Dersom minst tre av de fem har samme relasjon, er dette relasjonspreferansen til ledd A eller B.

Analogi: Fins det analoge sammensetninger som letter tolkninga av sammensetning X? Denne faktoren nevnes som vesentlig for en sammensetnings gjennomsiktighet av både Loenheim (2019) og Svanlund (2002). Førstnevnte konkluderer med at «huruvida en sammansättning som är ny för den enskilde språkbrukaren framstår som genomskinlig sammanhänger med vilken tillgång hon har till analogibaser» (2019:281). Her kan det i enkelte tilfeller være tilstrekkelig med én eneste analog sammensetning, som i tilfellet med *blåstrømpe* 'høyrevridd feminist', som trolig tol-

kes i analogi med *rødstrømpe* 'venstrevridd feminist', eller tilfellet *forkvinne* som analogi til *formann*. Begge disse tilfellene er imidlertid enveiskjørte i den forstand at den minst etablerte termen forstås i analogi med den mest etablerte, og ikke motsatt. Analoge sammensetninger kan i prinsippet identifiseres på mange nivåer, men jeg inntar her et restriktivt syn på analogi og medregner kun sammensetninger med en helt åpenbar mønster- eller betydningslikhet. For at sammensetning X skal være analog med sammensetning Y, må sammensetning X være mindre frekvent enn sammensetning Y. Videre må X og Y ha ledd A eller B til felles, og det resterende leddet må inngå i samme leksikalske felt. Alternativt må sammensetning X og Y ha tilsvarende definisjoner i enten BOB eller NAOB, samtidig som de har minst ett ledd til felles. *Fjellhare* er en analogi til *fjellrev* fordi den er mindre frekvent, fordi den har forleddet *fjell-* til felles, og fordi *hare* inngår i samme leksikalske felt som *rev*. **Framgangsmåte:** Sammensetning X får +1 i gjennomsiktighet hvis det fins en analog i LBK til sammensetning X etter disse kriteriene.

Motiveringsgrad og *relasjonsmønster* vektes høyere enn andre faktorer (her er det mulig å få opptil henholdsvis 3 og 2 poeng) siden disse trolig har størst innflytelse på gjennomsiktigheten til sammensetninger.

For alle flertydige sammensetninger blir den presumptivt minst gjennomsiktige betydninga bedømt. Det er nemlig denne som på semantisk grunnlag motiverer ordbokstatusen til en gitt sammensetning.

3.2. Data

I den påfølgende undersøkelsen, som viser modellen i bruk, vil jeg analysere gjennomsiktigheten til tre grupper sammensetninger, nemlig sammensetninger som begynner med henholdsvis *svart-*, *vandre-* og *tanke-*. Utvalget består av samtlige toledda sammenset-

ninger med disse forledda i BOB, komplementert med de 15 mest frekvente sammensetningene i LBK (se oversikt i tabell 1). Disse utvalgskriteriene skaper en skeivhet i utvalget i retning av etablerte sammensetninger. Ulempen oppveies imidlertid av at det alltid er det mest frekvente segmentet av sammensetninger med et gitt forledd som er aktuelt for ordbokoppføring, i det minste i arbeid med allmennordbøker, som har i oppgave å fange det mest sentrale ordforrådet i et språk.

4. Resultater

I dette kapitlet vil jeg kort presentere resultatene for de tre sammensetningsstrekkene, før jeg i neste kapittel går over til å diskutere modellens fordeler og ulemper. En oversikt over gjennomsiktighetsgraden til hver enkelt sammensetning beregnet etter modellen gis i tabell 1 mot slutten av kapitlet. I denne oversikten mangler kolonnen lengst til høyre, altså den mest gjennomsiktige enden av skalaen, siden ingen sammensetninger fikk maksimal uttelling på gjennomsiktighet.

4.1. *Svart-*

Svart-sammensetningene (heretter s-sammensetninger) fordeler seg over så godt som hele gjennomsiktighetsskalaen, fra de mest gjennomsiktige (til dømes *svartbrun* og *svarthåret*) til en større samling som er minimalt gjennomsiktige (til dømes *svartsjuk* og *svartemarje*). Bakgrunnen for spredninga er trolig de polyseme egenskapene til forleddet *svart-*, som medfører at ulike mentale forestillinger aktiveres i ulike sammensetninger. I materialet opptrer *svart-* med disse betydningene:

- 'som har svart farge', til dømes *svartjord*, *svartspett*, *svartovn*, *svarthåret*
- 'som kjennetegnes av noe med fargen svart', til dømes *svartedauden*, *svartemarje*, *svartkopp*, *svartskjorte*
- 'djevelsk' eller 'okkult', til dømes *svartemannen*, *svartebok*, *svartekunst*
- 'dyster' eller 'nedstemt', til dømes *svartsjuk*, *svartsyn*, *svartsinn*
- 'ulovlig' eller 'bannlyst', til dømes *svartebørs*, *svarteliste*

Kun 12 av 39 s-sammensetninger kan parafraseres mer eller mindre tilfredsstillende uten andre innholdsord. Til dømes er *svarthåret* og *svartovn* mer eller mindre ekvivalente med frasene *med svart hår* og *ovn som er svart*, mens *svartebørs* ikke kan erstattes av frasen *børs som er svart*.

Når det gjelder motiveringsgrad, samsvarte 19 av 39 sammensetninger med betydning 1 i BOB: 'med farge som sot eller kull'. Her fins bl.a. *svartand*, *svartbak* og *svartrot* som denoterer noe med slik farge, mens *svartebok*, *svartmale* (i overført betydning) og *svartemannen* ikke er motivert av den konvensjonelle betydninga av *svart*.

5 av 39 sammensetninger får uttelling på bakgrunn av kriteriet disambiguering. Her har vi sammensetningene *svarthvit*, *svartbrun*, *svartsmusket*, *svartsladd* og *svartlakkere*. Disse er disambiguert med bakgrunn i at etterledda understreker fargebetydninga til forleddet *svart*.

Ingen av s-sammensetningene har fått uttelling på bakgrunn av relasjonspreferansene til forleddet. Dette har sin rot i at de 5 mest frekvente sammensetningene i LBK, *svart-hvitt*, *svartebørs*, *svartkledd*, *svarthåret* og *svarteper*, ikke utviser noen entydig relasjonspreferanse for *svart-*.

7 sammensetninger har fått uttelling for analogier. Disse er *svartand* (analogi: *svarttrost*), *svartebok* (*hvitebok*), *svarthvit*[5] (*gråhvit*), *svarthåret* (*rødhåret*), *svartskjorte* (*brunskjorte*), *svartspett* (*svarttrost*) og *svartbrun* (*rødbrun*). Mest tvilsom er kanskje *svarthvit* (*gråhvit*), siden analogien som oftest betyr 'grålig hvit'. I og med at begge kan ha kopulativ betydning 'a og b', har *svarthvit* likevel fått uttelling for analogi.

Som helhet fordeler s-sammensetningene seg utover 7 av 8 trinn på gjennomsiktighetsskalaen. En viss overvekt av disse har lav gjennomsiktighet. De ikke-ordbokførte sammensetningene havner i den gjennomsiktige enden. Ellers finner man her også en overvekt av adjektiv-adjektiv- og adjektiv-verb-sammensetninger. I den ugjennomsiktige enden fins en del metaforiske og metonymiske sammensetninger. Til dømes er *svart-hvitt* i NAOB definert som 'unyansert'.

4.2. *Tanke-*

Tanke-sammensetningene (t-sammensetningene) konsentrerer seg mer mot sentrum av skalaen enn s-sammensetningene, med en liten overvekt i retning av ugjennomsiktighet. Forleddet *tanke-* har ikke distinkt ulike betydninger i t-sammensetningene, men noe betydningsvariasjon er det definitivt å spore. Det virker til dømes rimelig at *tankeretning* bygger på en annen forståelse av tankebegrepet enn *tankegymnastikk*. Førstnevnte har en ekvivalent med forleddet *idé-*, altså *idéretning*, mens *idégymnastikk* ikke er et godt synonym til *tankegymnastikk*.

8 av 30 sammensetninger lar seg parafrasere forholdsvis tilfredsstillende. Blant dem er typer med verbavledede etterledd, som *tankeoverføring* og *tankelesing*. Disse er parallelle med *overføring av tanker* og *lesing av tanker*. Videre kan *tankerekke* omsettes

5 Merk at *svarthvit* og *svart-hvitt* ikke er samme sammensetning. De har ulike oppslag i både BOB og NAOB.

til *rekke med tanker* og *tanketom* til *tom for tanker*. *Sprang i/med tanker* og *gods av tanker* er derimot lite tilfredsstillende erstatninger for *tankesprang* og *tankegods*.

Alle t-sammensetningene er motivert av betydning 1 av *tanke* i BOB, 'tankevirksomhet, forestilling i bevisstheten'.

Det er få t-sammensetninger der et av ledda disambiguerer det andre. De eneste tilfellene er *tankebygning* og *tankevekkende*, der forleddet *tanke-* peker ut betydninga til etterleddet.

Ingen av t-sammensetningene blir gradert som mer gjennomsiktige på bakgrunn av relasjonspreferansene til forleddet *tanke-*. De mest frekvente sammensetningene med *tanke-* i LBK er *tankegang*, *-vekkende*, *-gods*, *-rekke* og *-kors*. I førstnevnte tolker jeg *gang* som en nominalisering av verbet gå. Forleddet *tanke-* er dermed agens til dette verbet, mens forleddet i *tankevekkende* er patiens. De tre gjenstående, *tankegods*, *-rekke* og *-kors*, fanges ikke uten videre opp av Eieslands (2015) relasjonskategorier. Er til dømes et *tankekors* et kors *i* tankene, et kors *av* tanker eller et kors som er *fra* eller *tilhører* en tanke? Det er ikke åpenbart hva relasjonen er her, og jeg konkluderer derfor med at den jevne språkbruker ikke vil oppleve disse sammensetningene som mer gjennomsiktige på bakgrunn av en tydelig relasjonspreferanse for forleddet *tanke-*. Enkelte av sammensetningene, deriblant *tankeleser*, *tankeoverføring* og *tanketom*, gir uttelling for gjennomsiktighet på bakgrunn av relasjonspreferansen til etterleddet.

4 av 30 t-sammensetninger er mer gjennomsiktige på bakgrunn av analoge sammensetninger. Disse er *tankebane* (analogi: *tankegang*), *tankegymnastikk* (*tankeøvelse*), *tankemodell* (*tankesystem*) og *tankeverden* (*forestillingsverden*).

I motsetning til de ikke-ordbokoppførte s-sammensetningene plasserer ikke de tilsvarende t-sammensetningene seg nær den gjennomsiktige enden av skalaen. Om noe har de en svak tendens til ugjennomsiktighet, og de er såleis gode ordbokkandidater etter dette kriteriet.

4.3. *Vandre-*

Vandre-sammensetningene (v-sammensetningene) fordeler seg utover den midtre delen av gjennomsiktighetsskalaen. Forleddet *vandre-* uttrykker ulike nyanser i de forskjellige sammensetningene, men samtlige relaterer seg på et eller annet nivå til bevegelse eller forflytning. V-sammensetningene er langt mindre frekvente enn s- og t-sammensetningene. Den mest frekvente, *vandrehistorie*, har 72 treff i LBK. Til sammenlikning har de mest frekvente s- og t-sammensetningene *svartkledd* og *tankegang* henholdsvis 198 og 964 treff. I motsatt ende har den minst frekvente v-sammensetninga i undersøkelsen skarve 4 treff i LBK, mens tilsvarende tall for s- og t-sammensetningene er henholdsvis 29 og 18.

11 av 23 v-sammensetninger lar seg parafrasere og får dermed uttelling for gjennomsiktighet. Av disse kan *vandrefolk = folk som vandrer*, *vandrehall = hall til å vandre i* og *vandrestav = stav til å vandre med* nevnes. På motsatt side kan *vandremandag ≠ mandag som vandrer*, *vandrepokal ≠ pokal som vandrer* og *vandrehistorie ≠ historie som vandrer* nevnes. I de to sistnevnte trengs tilleggene *mellom vinnere* og *mellom folk* for å få fullgode parafraser.

Også 6 av 23 v-sammensetninger er motivert av betydning 1 av *vandre* i BOB, 'gå rolig'. Disse er *vandremandag*, *hall*, *safari*, *-stav*, *-sløyfe* og *-tur*.

Blant v-sammensetningene er det kun én som etter min tolkning oppviser noen form for intern disambiguering ledda imellom, nemlig den mest gjennomsiktige, *vandretur*. Her peker forleddet *vandre-* ut betydning 2 av *tur* i BOB, 'kortere eller lengre reise, ferd, utflukt'.

De mest frekvente sammensetningene i LBK med *vandre-* er *vandrehistorie*, *-falk*, *-utstilling*, *-år* og *-pokal*. Med unntak av i *vandreår* er etterleddet her agens til verbet *vandre*. Etterleddet spesifiserer altså hva det er som vandrer, og vi kan i henhold til kriteriene i modellen slå fast at relasjonspreferansen til forleddet *vandre* er

å stå i et verb–agens-forhold til etterleddet. 16 av de 23 v-sammensetningene følger dette relasjonsmønsteret. Blant de negative dømene finner vi *vandrehall* (som ikke er en hall som vandrer), *vandremandag* og *vandresafari*.

5 v-sammensetninger har fått gjennomsiktighetsuttelling for mer frekvente analoge sammensetninger. Disse er *vandrefigur* (*gjennomgangsfigur*), *vandrefolk* (*flyttefolk*), *vandremotiv* (*gjennomgangsmotiv*), *vandrefugl* (*trekkfugl*) og *vandretur* (*spasertur*, *gåtur*). Analogien består i at et av ledda er realisert med et synonym (*gåtur*), eller at definisjonen til sammensetninga som helhet tilsvarer definisjonen til v-sammensetninga, slik tilfellet er med *gjennomgangsfigur*.

Det er ingen av v-sammensetningene som plasserer seg i skalaens ytterpunkter. Snarere hoper de seg opp mot sentrum, men med en betydelig skeivhet i retning lav gjennomsiktighet. Dette skyldes at de fleste av v-sammensetningene ikke bygger på den konvensjonelle betydninga av *vandre*.

4.4. Sammendrag av resultatene

Resultatene fra analysen (se tabell 1) av et sammensetningsutvalg med forledda *svart-*, *tanke-* og *vandre-* viser at man for hver forleddstype kan forvente seg en viss gjennomsiktighetsvariasjon. Ingen sammensetninger får full uttelling for samtlige faktorer, og kolonnen helt til høyre på skalaen er derfor ikke inkludert i tabell 1. S-sammensetningene varierer klart mest når det gjelder gjennomsiktighet, siden disse plasserer seg utover 7 av 8 trinn på gjennomsiktighetsskalaen.

Minst gjennomsiktig					**Mest gjennomsiktig**	
svart-	**svart-**	**svart-**	**svart-**	**svart-**	**svart-**	**svart-**
s-e-dauden	s-bak	s-e-børs	s-e-bok	s-and	s-sladd	s-hvit
s-e-mannen	s-male	s-jord	s-sinn	s-hyll	s-smusket	s-håret
s-e-marje	s-e-per	s-e-kunst		s-or	s-brent	s-brun
s--hvitt	s-skjorte	s-kvist		s-rot	s-kledd	s-lakkere
s-katt		s-e-liste (v)		s-spett		
s-kopp		s-e-liste (s)		s-still		
s-sjuk		s-spett		s-trost		
s-sjuke		s-år		s-ovn		
s-syn						
	tanke-	**tanke-**	**tanke-**	**tanke-**	**tanke-**	
	t-flukt	t-bane	t-arbeid	t-gymna-stikk	t-leser	
	t-gods	t-korn	t-bygning	t-vekkende	t-over-føring	
	t-kors	t-lesing	t-eksperi-ment		t-rekke	
	t-retning	t-modell	t-gang		t-tom	
	t-stiller	t-strek	t-sprang			
		t-vekker	t-spredt			
		t-sett	t-verden			
		t-smie	t-virksom-het			
			t-mønster			
			t-prosess			
			t-spinn			
	vandre-	**vandre-**	**vandre-**	**vandre-**	**vandre-**	
	v-historie	v-figur	v-mandag	v-bibliotek	v-folk	
	v-pokal	v-motiv	v-sagn	v-falk	v-fugl	
	v-skjold	v-stjerne	v-arbeider	v-nyre	v-stav	
		v-sløyfe	v-due	v-utstilling	v-tur	
		v-år		v-hall		
				v-maur		
				v-safari		

Tabell 1: Gjennomsiktighetsgradering for sammensetninger med forledda *svart-*, *tanke-* og *vandre-*. Understrekede sammensetninger er ikke oppført i BOB. Markeringene *(v)* og *(s)* står henholdsvis for verb og substantiv.

5. Diskusjon

Gjennomsiktighetsmodellen som er presentert og anvendt her, gir en finkornet analyse av ulike sammensetninger og plasserer hver av dem på en skala. Ulike sammensetninger kan av ulike grunner bli oppfattet som gjennomsiktige, og det kan derfor være at vidt forskjellige sammensetningstyper, altså sammensetninger som har ulike kombinasjoner av ordklasser i ledda sine, får samme gjennomsiktighetsgradering, samtidig som sammensetninger av samme type kan få forskjellig gjennomsiktighetsgradering. Det er en udiskutabel styrke ved modellen at den for hver sammensetning uavhengig av type kan bedømme gjennomsiktigheten ut fra de samme faktorene. Om ulike sammensetningstyper har ulik tilbøyelighet til å være gjennomsiktige, vil dette bli fanget inn av faktorene som er innebygd i modellen, bl.a. tilleggselement og relasjonsmønster. Anvendeligheten på tvers av ordklasser er dessuten gunstig for leksikografer siden den muliggjør en arbeidsmåte der samme redaktør redigerer hele sammensetningsstrekk, som ofte kan inneholde ulike typer.

Det er liten grunn til å tro at gjennomsiktighet er systematisk fordelt på bakgrunn av ordklassetilhørigheten til ledda. Som tabell 1 viser, ligger adjektiv-adjektiv-sammensetningene *svartsjuk* og *svartbrun* i hver sin ende av gjennomsiktighetsskalaen.

Modellen tar også høyde for formlikhet mellom ulike sammensetninger. Mindre frekvente sammensetninger (til dømes *svartand*) kan forstås gjennom mer frekvente sammensetninger (til dømes *svartspett*), men ikke vice versa. Dette kan føre til at *svartand* blir bedømt som mer gjennomsiktig enn *svartspett*. Semantisk er dette rimelig, for det vil nettopp være flere faktorer som bidrar til å øke gjennomsiktigheten til *svartand* enn til *svartspett*. Det viser imidlertid hvor nødvendig det er å kombinere denne modellen med en distribusjonsanalyse som fanger opp det mest aktive ordforrådet, der *svartspett* trolig går foran *svartand*. Poenget

er uansett ikke å skille mellom sammensetninger som står rett ved siden av hverandre i tabellen, men snarere å undersøke omtrent hvor en gitt sammensetning havner.

Enkelte andre resultater kan framstå som kontraintuitive. Hvorfor er til dømes *svartbrun* blant de mest gjennomsiktige sammensetningene når den har en uklar relasjon mellom ledda og derfor er flertydig mellom betydningene 'svartaktig brun', 'brunaktig svart' og 'brun og svart'? Dette belyser muligens at modellen ikke tar nok høyde for ulike potensielle totalbetydninger. På den annen side kan man hevde at *svartbrun* er et godt døme på en sammensetning med liten leksikografisk relevans til tross for høy frekvens. Det at den er flertydig, kommer delvis av at den kan være både kopulativ og determinativ, men delvis også av lav konvensjonalitet, som dessuten kan tjene som en del av forklaringa på hvorfor ledda ikke har antatt noen spesialiserte betydninger, og hvorfor relasjonen mellom ledda er uklar.

De operasjonelle kriteriene i kapittel 3 gir langt på vei klare instrukser for hvordan en gitt sammensetning rent praktisk skal vektes i henhold til gjennomsiktighetsfaktorene. Faktoren relasjonsmønster er vanskeligst å operasjonalisere siden den krever distinkte relasjonspreferanser for både for- og etterleddet i hver enkelt sammensetning. I mange tilfeller er det tidkrevende å utlede disse. En mulig løsning er å operere med relasjonspreferanser med mest mulig grovmaskede kategorier.

En annen ulempe er at analysen tar utgangspunkt i ordbøker som premissleverandør for hva som er den konvensjonelle betydninga av et ord. Det er ingen garanti for at den øverste betydninga til et oppslagsord er den som oppleves som mest konvensjonell av et flertall språkbrukere. Definisjoner og betydningsrekkefølge kan være arvet fra et tidligere ordbokverk som oppstod i en helt annen språklig kontekst. Dette kan lede til gale antakelser om konvensjonalitet, som igjen leder til at man får upresise beregninger av sammensetningers motiveringsgrad.

En annen mulig svakhet ved gjennomsiktighetsmodellen er at den ikke kan anvendes mekanisk uten at leksikografens språkkunnskaper kommer i spill. Dette er til en viss grad tilfellet for alle faktorene. Ulike leksikografer kan dermed tenkes å ende opp med ulike resultater når de bruker modellen. Det er imidlertid ikke sikkert at denne variasjonen er problematisk. Når det kommer til stykket, er det nettopp dette som, i alle fall tradisjonelt, kjennetegner ordbøker – at de er sammensatt og kvalitetsvurdert av ulike redaktører med mer eller mindre ulike språklige bakgrunner og formeninger om ords betydning. At ulike redaktører kommer til ulike resultater, er en del av det som forhindrer leksikografisk ensretting, og som sørger for at mer enn ett perspektiv er representert i ordboka.

Gjennomsiktighetsmodellen er utformet i tråd med et syn på ordbøker som hevder leksikografenes kvalitetsvurderinger som vesentlige og nødvendige. Målet er dermed ikke at gjennomsiktigheten i sammensetninger skal kunne regnes ut mekanisk uten språklige vurderinger, men snarere at gjennomsiktighetsmodellen kan gi leksikografen konkrete holdepunkter for hvilke aspekter ved sammensetninger som er relevante å trekke inn når man bedømmer gjennomsiktighet.

6. Oppsummering og videre forskning

Den mest pålitelige evalueringa av gjennomsiktighetsmodellen er trolig et fortolkningseksperiment der man kan observere empirisk om språkbrukeres vurderinger er i tråd med modellens påstander om gjennomsiktighet. Bare med en slik metode kan man verifisere om modellens faktorer faktisk påvirker gjennomsiktighetsgraden, og avdekke om det eventuelt fins andre faktorer som er i spill.

Videre må man supplere gjennomsiktighetsmodellen med undersøkelser av den enkelte sammensetnings utbredelse på tvers av tekstsjangre, kilder og tid.

Litteratur

Ordbøker

BOB = *Bokmålsordboka*. Bergen: Universitetet i Bergen og Språkrådet. <http://ordbok.uib.no> (august 2019).

NAOB = Tor Guttu mfl. (red.): *Det Norske Akademis ordbok*. Oslo: Det Norske Akademi for Språk og Litteratur. <https://www.naob.no> (august 2019).

NOB = *Nynorskordboka.* Bergen: Universitetet i Bergen og Språkrådet. <http://ordbok.uib.no> (august 2019).

Annen litteratur

Bakken, Kristin (1998): *Leksikalisering av sammensetninger: en studie av leksikaliseringsprosessen belyst ved et gammelnorsk diplommateriale fra 1300-tallet*. Det historisk-filosofiske fakultet, Universitetet i Oslo.

Eiesland, Eli Anne (2015): *The semantics of Norwegian noun-noun compounds: a corpus-based study*. Det humanistiske fakultet, Universitetet i Oslo.

Fjeld, Ruth V. & Lars S. Vikør (2008): *Ord og ordbøker: ei innføring i leksikologi og leksikografi.* Kristiansand: Høyskoleforlaget.

Kjelsvik, Bjørghild (2018): *Retningslinjer for sms i BOB og NOB*. Upublisert internt notat for Revisjonsprosjektet.

Knudsen, Rune L. & Ruth V. Fjeld (2013): LBK2013: A Balanced, Annotated National Corpus for Norwegian Bokmål. I: *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013*. *NEALT Proceedings Series* 19, 12–20.

LBK = Leksikografisk bokmålskorpus. Tekstlaboratoriet, Institutt for lingvistiske og nordiske studier, Universitetet i Oslo. <http://www.hf.uio.no/iln/tjenester/kunnskap/samlinger/bokmal/veiledningkorpus/> (august 2019).

Loenheim, Lisa (2019): *Att tolka det sammansatta. Befästning och mönster i första- och andraspråkstalares tolkning av sammansättningar*. Meijerbergs arkiv för svensk ordforskning 43, Göteborgs universitet.

Maguire, Phil, Rebecca Maguire & Arthur W.S. Cater (2010): The influence of interactional semantic patterns on the interpretation of noun–noun compounds. I: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36:2, 288–297.

Ryder, Mary (1994): *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*. University of California Publications in Linguistics 123. Berkeley: University of California Press.

Schmid, Hans-Jörg (2015): A blueprint of the entrenchment-and-conventionalization model. I: *Yearbook of the German Cognitive Linguistics Association*, 3:1, 3–26.

Shoben, Edward J. & Christina L. Gagné (1997): Thematic relations and the creation of combined concepts. I: Thomas B. Ward, Steven M. Smith & Jyotsna Ed Vaid (eds.): *Creative thought: An investigation of conceptual structures and processes*. American Psychological Association, 31–50.

Svanlund, Jan (2002): Lexikalisering. I: *Språk och stil* 12, 7–45.

Svanlund, Jan (2009): *Lexikal etablering: En korpusundersökning av hur nya sammansättningar konventionaliseras och får sin betydelse*. Acta Universitatis Stockholmiensis.

Mikkel Ekeland Paulsen
doktorgradsstipendiat i nordisk
Institutt for lingvistiske, litterære og estetiske studier
Universitetet i Bergen
Haakon Shetteligs plass 7
NO-5007 Bergen
Mikkel.Paulsen@uib.no

III Delstudie 3: Wheat or chaff? A Compound Selection Model Based on Look-Up Data

International Journal Of Lexicography, 2023, **XX**, 1–19
https://doi.org/10.1093/ijl/ecad013
Advance access publication 30 May 2023
Article

Wheat or Chaff? A Compound Selection Model Based on Look-Up Data

Mikkel Ekeland Paulsen

University of Bergen, Norway Norway
(Mikkel.Paulsen@uib.no)

Abstract

Which compounds should be included in general-purpose dictionaries is often an open question that is answered with a case-by-case consideration of all compounds above a certain corpus frequency threshold. Another way to determine which compounds should be listed, is to examine which compounds, or rather which compound properties, are in demand by the users. This study uses look-up data from the two officially sanctioned, general-purpose dictionaries of Norwegian (Bokmålsordboka and Nynorskordboka) to derive an explicit compound selection model that performs with comparable sensitivity and specificity as the traditional procedure. These findings demonstrate that it is indeed possible to arrive at a fully operational and explicit compound selection model that meets the needs of users. With such a tool at their disposal, lexicographers would be able to separate the wheat from the chaff in the boundless field that is the compound lexicon of North Germanic Languages.

Keywords: compound selection, Norwegian, look-up data, corpus data, conditional inference trees

1. Background

Selecting compounds for general-purpose dictionaries of any North Germanic language resembles the act of separating the wheat from the chaff, aside from the fact that in the compound selection scenario there are no physical properties that disclose which compounds correspond to the respective roles. Instead, lexicographers must base their decisions on linguistic and distributional properties of each compound. One way to determine which properties warrant inclusion is to observe which properties are in demand. This study aims to identify the variables that govern the look-up interest into compound lemmas in online dictionaries.

A particular challenge that pertains to compound selection in these languages is the ubiquity and productivity of compounds in language use. As an example, in the medium-sized corpus *Leksikografisk bokmålskorpus* (henceforth LBK), (Fjeld et al. 2020) the string that makes up the semi-frequent word *maskin* 'machine' enters into approximately 2,000 different compound types. Obviously, not all of these compounds can or should go into a general-purpose dictionary. Lexicographers therefore need rigorous methods and criteria for compound selection to be able to extract a selection of compounds that serves the needs of dictionary users. Although there can never be a gold standard for what makes the ideal lemmalist of compounds across dictionaries with different scopes and purposes, look-up behaviour in an online dictionary may give insight into the interests and needs of the users of that particular dictionary. This user interest should at least be a part of the equation in the assessment of the

lemmalist of a given dictionary. We may therefore view look-up statistics as an important empirical foundation for both assessment and enhancement of an existing lemmalist.

Corpus frequency is normally understood as a numerical value that reflects how many occurrences there are of a corpus item relative to the size of the corpus. In an online Swahili-English dictionary, De Schryver et al. (2006) found a positive correlation between look-up and corpus frequency among the 5000 most frequent words in the corpus. Trap-Jensen et al. (2014) suggest that corpus frequency is an important predictor of look-up frequency in a monolingual Danish dictionary for about the 20 000 most frequent lemmas in a balanced corpus. The findings of Wolfer et al. (2014) and Müller-Spitzer et al. (2015) also indicate that corpus frequency is an important predictor of look-up frequency in monolingual German dictionaries. This finding is confirmed by De Schryver et al. (2019) in a further study on the same online Swahili-English dictionary. Although one might assume that there are important differences between the look-up behaviour in monolingual and bilingual dictionaries, corpus frequency seems to be an important predictor of look-up frequency in both dictionary types.

A weakness of the above-mentioned studies is that their quantitative focus is limited to corpus frequency, which is an unreliable and somewhat invalid measure of frequency of occurrence, especially if frequency of occurrence is perceived as a token of overall word importance or commonness (Gries 2008: 404, Paulsen 2022). For one thing, it suffices for a corpus item to be very frequent in one limited part of the corpus in order to seem frequent in the corpus as a whole. For example, if the corpus contains a book on goatfish, then the word *goatfish* would occur surprisingly frequently in the corpus as a whole. Goatfish-effects make frequency unreliable in that different language samples will generate very different frequencies for such words. If the frequency of a word is completely different for every language sample, one should be cautious about inferring from frequency estimates to properties of the language variety that the corpus is sampled from.

A way to escape arbitrary goatfish-effects is to include a measure of dispersion that evaluates the degree to which the occurrences of a corpus item are evenly spread throughout the corpus. A word which is both well-dispersed and frequent in the corpus sample is more likely to be commonly occurring in the language variety that the corpus is sampled from than a frequent word whose occurrences are clustered together. Put briefly, dispersion influences what inferences we can make from the frequency score (Paulsen 2022).

Frequency and dispersion, then, are quantitative variables that with all likelihood have a bearing on the look-up frequency of compounds. Since dictionaries are consulted both in productive and receptive contexts, e.g., both when writing and reading, it is reasonable to assume that words which are frequently read or written will typically be looked up more often than words which are seldom read or written.

The present study will also survey a number of **qualitative** variables that may affect which Norwegian compounds users look for, namely semantic transparency, part of speech (POS), whether or not the compound has an interfix, the number of spelling variants of a compound, and one that is particular to parallel dictionaries: whether or not there is equivalence between the two Norwegian written languages *Bokmål* and *Nynorsk*.[1],[2]

A satisfactory model for compound inclusion must integrate considerations pertaining to both linguistic properties and usage. In addition, the criteria and variables on which the model operates must be explicit and reproducible. The major aim of this study is to propose such a model for compound selection in the officially sanctioned general-purpose dictionaries of Norwegian, *Bokmålsordboka* (henceforth *BOB*) and *Nynorskordboka* (henceforth *NOB*) (and, by extension, of other languages that are morphologically similar in terms of compounding). In order to develop such a model, the following steps are taken:

- Based on a sample of compounds, the implicit standard model (henceforth *StandMod*) that has served to create the current lemmalist of compounds in the BOB is examined with respect to its association with the variables in the study.

- Through a mixed set of methods, the variables in the study are investigated with regard to their association with look-up statistics.
- Based on the results of the investigation in the previous step, an alternative model comprised of explicit criteria is developed (henceforth *The Look-Up Predictor Model*, abbreviated *LookMod*).
- Using the look-up data, the performance of LookMod is evaluated and compared with the performance of StandMod.

Historically, it is unknown which set of criteria has prompted the current compound list in the BOB. In other words, it is not explicitly stated what reasoning lies behind the inclusion of a given compound and it is likely that different editors have partly based their decisions on their individual and idiosyncratic intuition.

The findings of the above-mentioned procedure in part validate the StandMod by demonstrating that its output, the current lemmalist of the BOB, is to a large extent in harmony with the look-up interests of the dictionary users. Using variable levels deducted from look-up statistics, the LookMod performs at nearly the same level of specificity and sensitivity as StandMod. Although these are promising results for LookMod, it nevertheless shows that performance alone does not motivate a shift in lexicographical practice. However, the LookMod has some other advantages. Firstly, it is more transparent than the StandMod. Secondly, it is more objective and less reliant on the intuition of a given lexicographer. And thirdly, it could potentially be less time-consuming than the StandMod overall, if one could find a way to automatically annotate all compounds in a corpus according to (most of) the variables in the model. It should be noted that this is somewhat far fetched currently, since we still lack a procedure for automatic detection of compounds.

Although the advantages of the LookMod are not trivial, its primary usefulness is the information that it conveys about the variables it contains (and ignore). First and foremost, corpus dispersion stands out as an important predictor of look-up interest. In other words, dispersion is a variable that ought to be included in whatever model lexicographers apply.

2. Norwegian compounds

For the purpose of this study, a compound is defined as a lexeme whose stem is comprised of two individual stems that are both found as stems of separately occurring lexemes. [3] These stems may be either compounds, derivatives or root words. Borrowed compounds (e.g., *airbag*) are excluded from this definition unless both constituents are stems of separately occurring lexemes, in this study operationalised as listed in the BOB, while borderline cases between derivation and compounding are generally accepted as compounds.

Norwegian compounds are generally right-headed, which means that the second constituent is the semantic and grammatical head, whereas the first constituent functions as a modifier. Norwegian compounds may also contain an interfix that is added as a suffix on the first constituent, as in *fortauskant* 'pavement' + $[s]_{interfix}$ + 'edge'. Generally speaking, interfix inclusion is a property of the modifier stem. For instance, *fortau* 'pavement' is consistently suffixed, whereas e.g., *vind* 'wind' is never suffixed. There are however many examples of stems that are variably suffixed (see Kulbrandstad & Kinn 2016 for examples).

Compounds are ubiquitous in Norwegian. This fact compels lexicographers to pick and choose from an unbounded list of candidates for dictionary inclusion. The productivity of Norwegian compounds also extends to grammatical variability: Nominals, adjectives, verbs, prepositions and adverbs are all fairly productive as compound constituents. Still there is no doubt that noun-noun compounds are the most productive type (see also Section 3.1).

Fjeld & Vikør (2008) argue that 'semantic transparency' is an important factor in a compound selecting scheme. There is no consensus on the exact meaning of 'semantic transparency', but it roughly refers to whether the meaning of the compound can be inferred from the meaning of its parts (Schäfer (2018) gives an overview over vastly different definitions

of semantic transparency). A simple operationalisation of semantic transparency that is utilised in this study is *degree of motivation* as it is defined by Svanlund (2002).

3. Data and method

This section outlines the dataset and the variables of the study.

3.1. Compound sample

The analyses of the current study will be performed on a sample of compounds. The sample is collected with the following steps:

1. Each compound in the sample belongs to one of the five alphabetical stretches (henceforth segments) afrikaans – -aktig, bryllup – bukt, dverg – dørk, einstøing – eksterritorialrett and forstokka – forårsake. These are segments which have recently received a full revision in the BOB, each by a different editor. The BOB contains 1560 lemmas from these segments, 802 of which are compounds.
2. From these segments, any given compound is included in the sample if it fulfils at least one of the following criteria:
   a. It has an entry in the BOB.
   b. It has a minimum of 20 occurrences in the corpus LBK, which equals approx. 0.2 occurrences per million words (pmw).
   c. It occurs in the look-up data (see Section 3.2).

The criteria a, b and c yield 802, 570 and 919 compounds, respectively. There is substantial overlap between the groups and the unique contribution from each category amount to 112, 153 and 196 respectively. The complete sample amounts to 1206 compounds and contains both frequent and infrequent items that show a presumed typical spread across different parts of speech for both constituents. See Supplementary Material Online for tables 1 and 2 that display examples of compounds from different frequency bands in the LBK and the distribution of different compound types.

In the sample, nouns are dominant in both the modifier and head position. Noun-noun compounds account for approximately 75% of the sample, which probably reflects the linguistic reality that Norwegian lexicographers deal with, i.e., that the noun-noun pattern is far more productive than any other compound pattern. As a reference, 84% of the lexeme entries[4] in the BOB are nouns, and about 78% of all entries in the *Norwegian Academy Dictionary* are nouns. This nominal predominance indicates that part of speech needs to be controlled for in statistical analyses.

3.2. Look-up data

Every look-up in the BOB and the NOB is saved in a search log. Both dictionaries are freely accessible from the same interface, currently at ordbokene.no[5] and the look-up data have been obtained through analysis of log files generated from both dictionary

Table 1: Counts and proportions of BOB status at levels of look-up regularity.

look-up regularity	0		1-10		10-100		>100		total
In BOB	count	prop %	count	prop %	count	prop %	count	prop %	
0	153	53	151	46	83	23	16	7	403
1	134	47	178	54	285	77	206	93	803
total	287		329		368		222		1206

look-up and access through general-purpose search engines. With this approach, an inventory of search queries and their number of occurrences within a certain time frame has been generated.[6]

The look-up data are accessible through a webpage[7] that contains a selection of files with lists of all queries that have been performed more than a certain number of times in a given year. For this study, every query that has been performed at least 10 times inside one of the calendar years 2016 – 2020 is included in the look-up data. This means that the data contain information about the accumulated number of look-ups of each compound in the sample for each of the five years in which there are 10 or more look-ups.[8]

The look-up frequency data are not lemmatised and there is no way of knowing for certain which dictionary entry a user is seeking when they use for example the query expression *fortelt*. One could assume that they are after the compound *fortelt* lit. 'front tent', but it is not at all unlikely that they have misspelled the perfect participle form *fortalt* of the frequent verb *fortelle* 'tell' (not a compound). Not knowing the users' intentions, we cannot know what they are seeking in all instances. This fact makes the following operationalisations inescapable: 1) We assume that users do not misspell their queries, and 2) we count certain queries as a look-up for more than one entry. This way, queries which are homographical with the compounds in the sample and inflectional variants thereof count in the summation of the overall number of look-ups for those compounds. Punctuation and frontal or final white space in the queries are ignored. This means that the above example will count as a look-up of *fortelt* and not *fortelle*.

Some dictionary entries have multiple spelling variants, for instance are *dypsnø* and *djupsnø* 'deep snow' parallel headwords of the same entry. Such spelling variants are treated as members of the same lexeme. A further technicality relates to the fact that most queries at *ordbok.uib.no* return values from both BOB and NOB. Since querying into both dictionaries simultaneously is the default when people first enter the website either directly or via a search engine, the parallel look-ups are by far the most frequent, and the look-up data are therefore based on these parallel look-ups. For this reason, query expressions that match compounds in *Nynorsk* that are equivalent to the sampled compound lemmas also enter into the overall number of look-ups for their equivalents. This means that for example both the query *forståelsesfull* (which matches a compound in Bokmål meaning 'understanding' (adjective)) and *forståingsfull* (which matches the Nynorsk equivalent of the compound *forståelsesfull*) add to the look-ups of the same compound lemma.

3.3. Configuration of look-up variables

Four look-up variables are employed in this study.

- **Number of look-ups** is the accumulated number of look-up events that match a compound lexeme.

Table 2: Performance of both models on the original dataset up regularity

look-up regularity		0		1-10		10-100		>100		
		count	prop %	count	prop %	count	prop %	count	prop %	Sum
StandMod	0	153	53	151	46	83	23	16	7	403
	1	134	47	178	54	285	77	206	93	803
LookMod	0	126	44	156	47	102	28	6	3	390
	1	161	56	173	53	266	72	216	97	816

- **Look-up frequency** is the number of look-up events for a given compound lexeme divided by the total number of look-ups.
- **Look-up dispersion** reflects the distribution of look-up events matching a compound lexeme over look-up year and is operationalised as the average of a *Deviation of Proportions* estimation (as described by Gries 2008, henceforth *DP*) and a *Juilland's D* estimation (as described by Juilland et. al. 1971).[9] In other words, the look-up dispersion reflects the degree to which the look-up frequency for a compound lexeme is stable over the five years that the look-up data represent. In short, dispersion is high when the look-up frequency remains stable over time.
- **Look-up regularity** is the product of number of look-ups and look-up dispersion. This variable reflects number of look-ups while controlling for look-up dispersion. Since none of the compound lexemes in the dataset exhibits optimal dispersion, the look-up regularity scores are slightly lower than number of look-ups for each compound. Figure 1 shows the distribution of log10-scaled look-up regularity. The purpose of the scaling is to make the x-scale of the histogram easier to read, but note that 287 compounds with zero look-ups are left out of the figure. The histogram shows that many compounds lie in the 1-10 and 10-100 ranges, and that the number of items decreases gradually toward a look-up regularity of 1000.

The look-up variables serve as response variables in Sections 5 and 6. There are a number of predictors included in the study which are all listed and briefly explained below. For a detailed account of the distribution and configuration of the predictors, please see Supplementary Material Online.

- **Number of occurrences in corpus (NO)**: The absolute number of occurrences of a compound lexeme in the LBK.
- **Corpus dispersion (disp)**: The average dispersion of a compound lexeme in the LBK based on a DP- and Juilland's D-estimation of the domain- and yearwise distribution of lemma.
- **Degree of motivation (DoM)**: The degree to which the conventional denotations of the constituents are active in the denotation of the compound as a whole, 0 = No motivation, 1 = Motivated modifier, 2 = Motivated head, 3 = Fully motivated.
- POS of modifier and head (**POS_m** and **POS_h**).
- **Interfix**: Whether the compound has an interfix.
- **Parallelism (paral):** The degree to which a compound in Bokmål has an equivalent in Nynorsk, *No* = No obvious equivalent, *Partial* = Semantic equivalence without homography, *Full* = Semantic and homographic equivalence

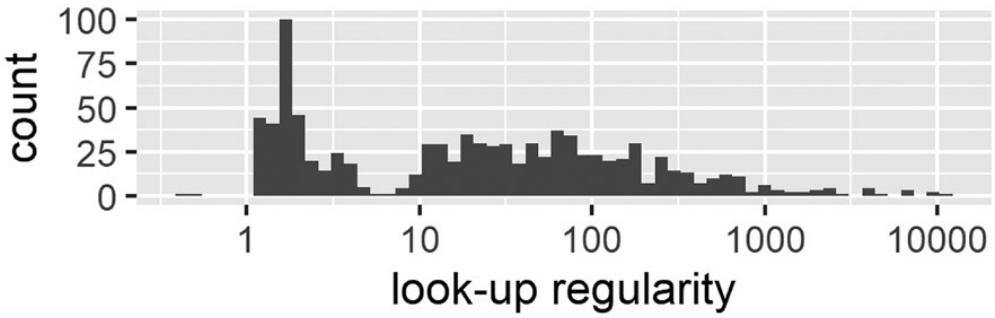


Figure 1: Histogram of logged look-up regularity in sample.

- **Number of spelling variants** (**Nvar**): Whether the compound has multiple spelling variants.
- **Salience:** The degree to which a compound deviates from the statistically most common properties with respect to the variables POS_m, POS_h, DoM, Nvar and paral.

Now that we are familiar with the data and variables of the study, we may proceed with an investigation of StandMod.

4. StandMod

In this chapter I present the background and broad tendencies of the current compound lemmalist in the BOB within the five segments that are investigated in this study. This list has emerged through a variety of methods and may be seen as the returned value of the StandMod. After an exploration of some of the characteristics of this model, I will evaluate its performance based on the degree to which it reflects the look-up interests of the dictionary users.

4.1. Background and broad tendencies

The set of compounds that are currently listed (*In BOB* = 1) in the BOB has come about through 12 years of compilation before the first publication, and 35 subsequent years of sporadic revisions and updates, the most recent in 2019-2020. During this period, the lexicographical practice has become increasingly corpus-based, and one might assume that the empirical foundation for lemma selection has been gradually strengthened. Another change is that the dictionary has moved from a printed to a digital format. Intuitively, one would assume that this change facilitates the listing of a greater number of compound lemmas since space is much less sparse in digital dictionaries compared to printed ones. The StandMod is nevertheless not explicitly formulated, and it is not known exactly which criteria the various lexicographers that have edited the dictionary have employed. All that is known is that many considerations underlie the selection of compound entries, and a given entry may be justified by for example corpus frequency, grammatical or semantic properties, or internal systematicity within the BOB or between the BOB and the NOB.

Figure 2 displays the relationship between log10-scaled NO (in the LBK) and BOB status. The inter-quartal ranges and the mean values (indicated by the dots) indicate that listed lemmas are associated with higher NO. However, there is considerable variation in the NO-values of both listed and unlisted compounds. In fact, unlisted compounds have a higher median NO than listed ones. This demonstrates that NO is not the only variable that governs dictionary inclusion.

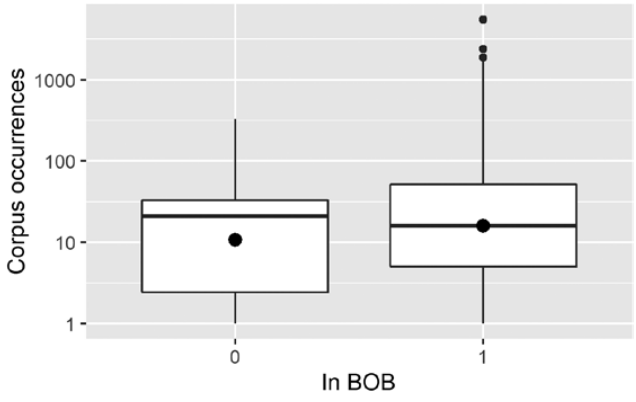


Figure 2: Distribution of NO over BOB status.

Furthermore, unlisted compounds are overrepresented among compounds with NO = 0 and 10–50, whereas listed compounds are overrepresented in the other categories (including 1–10) in Figure 3. In other words, NO seems to have a strong influence on dictionary inclusion when it exceeds 50, but there is substantial and seemingly random variation below this point. This indicates that high frequency may be an important qualifying factor for dictionary inclusion, especially beyond 400, but low frequency is evidently not systematically employed as a disqualifying criterion in StandMod.

There is substantial internal variation with respect to the dispersion score among listed and unlisted compounds, see Figure 4. The median disp value is higher for listed than unlisted items, but many of the listed items have a very low dispersion score.

More enlightening is perhaps Figure 5, which shows the proportions and absolute numbers of listed and unlisted compounds in different dispersion bands. The horizontal line indicates

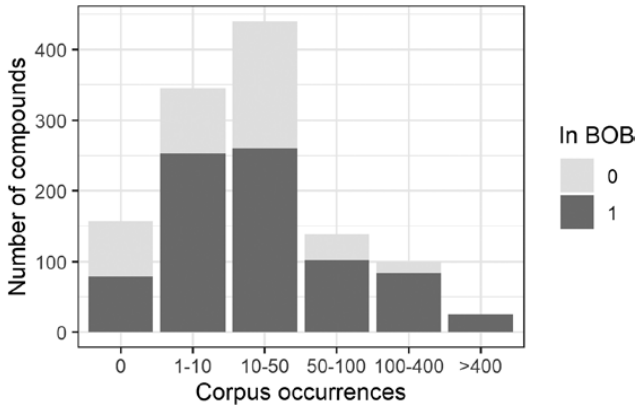


Figure 3: BOB status according to NO bands.

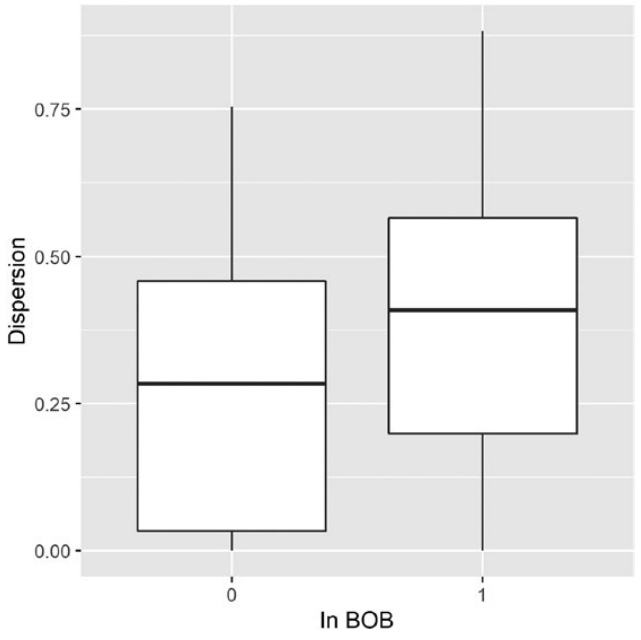


Figure 4: Distribution of disp over BOB status.

the global proportion of listed items. Here, we see a gradual increase in the proportion of listed items as the dispersion value increases. However, it is only in the upper bands 0.5-0.7 and >0.7 that listed lemmas are overrepresented compared to the percentage of listed items in the dataset (66.6%). While higher dispersion increases the likelihood of a compound being listed, low dispersion is evidently not employed as a disqualifying criterion in StandMod.

Deviation from full motivation is associated with an increase in the likelihood that a compound is listed, see Figure 6. Listed compound lemmas are overrepresented for all levels of Degree of Motivation below 3. It should, however, be noted that over 80% of the

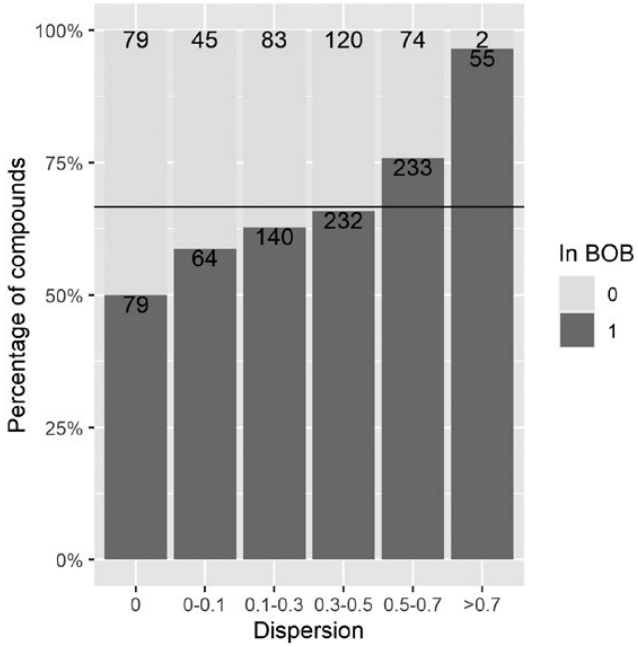


Figure 5: BOB status according to disp bands.

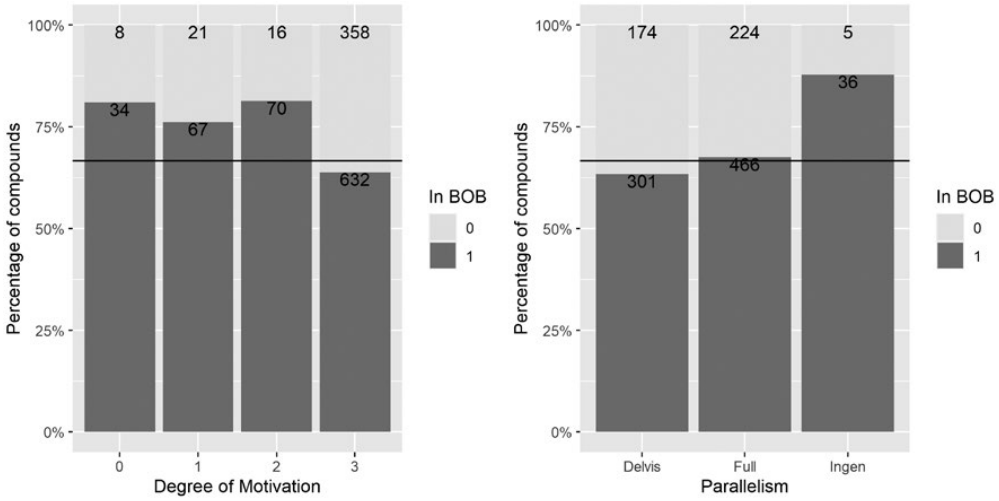


Figure 6: BOB status according to Degree of Motivation and Parallelism.

compounds in the dataset are fully motivated, so the distributions in the levels 0-2 are based on a small proportion of the compounds in the study.

Listed compounds are overrepresented among compounds in Bokmål with no obvious equivalent in Nynorsk, see Figure 6. Only 4% of the compounds in the study fall into this category, which indicates that the role of this variable is at best peripheral in StandMod. Note that the risk of influence from random variation increases in categories with few items. This is especially relevant among compounds with low degree of motivation and no parallelism.

Listed compounds are overrepresented among the adjectival, adverbial, interjectional, prepositional and verbal modifiers, which means that it is only among the compounds with nominal modifiers that listed compounds are underrepresented, see Figure 7. This indicates that StandMod includes a larger proportion of compounds with non-nominal modifiers than with nominal modifiers, presumably because the former are much less productive constructions in Norwegian. Since nominal compounds are much more productive than any other compound construction, one would expect this to be the chief construction among novel compounds also, which in turn would make nominals overrepresented among the unlisted compounds in the sample. The rather small effect of this can be seen on the left side of Figure 7.

A similar pattern can be found with respect to nominal heads. The plot on the right in Figure 7 indicates that listed compounds are overrepresented in all categories except for the nominal one, where they are underrepresented. It should, however, be noted that the distribution of BOB status in compounds with nominal heads is more or less identical with the global distribution in the sample (65%).

Number of spelling variants (Nvar) and interfix status do not demonstrate any particular tendency towards association with BOB status.

We may conclude that StandMod has a certain degree of covariation with the distributional variables Number of occurrences (NO) and dispersion (disp), and the semantic variable Degree of Motivation (DoM), and that a vast majority of the unlisted compounds in the dataset have nominal modifiers and/or heads. A generalised linear regression approach has also been attempted, but it has not helped indicate any further characteristics of StandMod that are not visible from the graphs presented above.[10] There is, in other words, substantial variation in BOB status that the variables in this study cannot account for. We

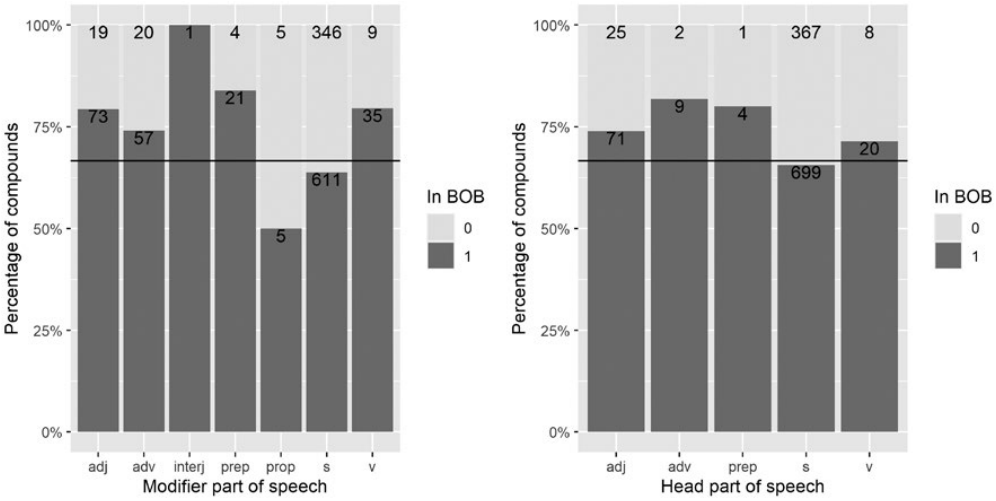


Figure 7: BOB status according to POS_m and POS_h.

may therefore conclude that a defining trait of StandMod is that it is multifaceted and flexible in nature, and that there are grounds for including a compound beyond the variables included here.

4.2. Evaluating StandMod

A way to evaluate the lemmalist of the current version of the BOB, and thereby the model that has produced it, is to investigate the extent to which the lemmalist is in harmony with the interests of the dictionary users. The results of such an investigation are presented in the following, using look-up statistics as an operationalisation of user interest.

A starting point for the evaluation is a simple analysis of the two possible sources of discordance between the dictionary and the look-up statistics, namely *unvisited entries* and *lacunas* (i.e., unlisted lemmas that are looked up fairly regularly) which indicate the specificity and sensitivity, respectively. In practice, unvisited entries are often quite unproblematic. There might be systemic reasons to include certain compounds, regardless of their interest to the dictionary users. However, a large number of unvisited entries coupled with a large number of lacunas might indicate that there is a lack of harmony between the selectional criteria of the dictionary and the needs of the users. The percentage of visited entries relative to unvisited ones is thus worth some scrutiny.

It is not obvious what the appropriate threshold for a lacuna is, but it would be unreasonable to treat every unlisted and looked-up compound as a lacuna. There might for instance be systemic and language-specific reasons for not listing a compound even though it is looked up, e.g., people might look up foreign words.

Table 1 gives an overview of the numbers of listed and unlisted compounds in the sample contingent on look-up regularity. Slightly fewer than 50% of the compounds with a look-up regularity of 0 are listed, while the same number for compounds that are looked up a few times is 54%. Further, the number and proportion of listed compounds increase as the look-up regularity increases beyond 10. Among the most looked up compounds, almost all compounds are listed.

If we turn to the two sources of discordance, there are 134 unvisited entries, which means that approx. 83% of the listed lemmas in the sample are visited often enough to appear in the look-up statistics. Furthermore, 151 + 83 + 16=250 lemmas, i.e., 62% of the unlisted lemmas in the sample, are looked up and are therefore potential lacunas. A majority of these are not obvious lacunas since they reside in the 1–10 look-up regularity range, but 99 of them have a look-up regularity >10, of which 16 have a look-up regularity >100. Based on look-up regularity alone, there appears to be room for improvement of both the specificity and the sensitivity of StandMod. To ascertain whether the unlisted lemmas are in fact lacunas, we can inspect the 16 unlisted compounds with a look-up regularity exceeding 100.

> *ajourholde* 'keep up to date', *bråvåkne* 'wake suddenly', *budrunde* 'bidding round', *boforhold* 'living condition', *dybdeintervju* 'in-depth interview', *dybdelæring* 'in-depth learning', *dyptpløyende* 'that plow deep', *dødslengsel* 'lit. death longing', *dødsspiral* 'death spiral', *dømesvis* 'for example', *fortauskant* 'curb' (lit. 'pavement edge'), *forutberegnelig* 'predictable' (lit. 'precalculable'), *forutenom* 'besides' (preposition), *forutgå* 'precede' (lit. 'pre go'), *forutsaker**

Of these, only *forutsaker* is a questionable candidate since it is most probably a misspelling, although it is not obvious which word it is a misspelling of. Since the remaining 15 lemmas are perfectly acceptable dictionary entries, we may, on the basis of look-up regularity, conclude that there are at least 15 unlisted lemmas that the BOB could benefit from including. If we set the bar at minimum 10 in look-up regularity, StandMod results in between 15 and 99 lacunas, i.e., 2.5 - 17% of the compounds with look-up regularity > 10. Whether or not this is an acceptably low proportion of lacunas is hard to judge without having done the

same calculation on other general-purpose dictionaries. But it is clear enough that there is at least some room for improving StandMod. In the following chapter, I will explore predictors of user interest in order to develop an alternative model.

5. An alternative model

Since StandMod results from multiple lexicographers' practices and a certain portion of its variables remains unknown (see Section 1), the best way to achieve further improvements with respect to the level of lacunas in StandMod is to first replace it with a model that can be formulated explicitly. Therefore, this chapter will proceed with an attempt to find predictors of user interest. From this I will derive a model that specifies 1) variables to consider in lexicographic selection, 2) conditions under which to consider each variable, and 3) sensible cut-off points for the selected variables in given circumstances. Finally, the derived model will be evaluated through lacuna analyses of its performance on both the data from which it was derived and an independent set of test data.

5.1. Conditional inference trees and random forests

A starting point for deriving such a model is to use conditional inference trees (henceforth *cits*) with look-up regularity as the response variable to illustrate how the variables included in this study may operate together to identify groupings of compounds that are frequently looked up, see Supplementary Material Online, Strobl et al. (2009), Tagliamonte & Baayen (2012) or Levshina (2015) for a detailed description of this method.

Cits are useful because they have the capacity to capture complex interactions between predictors (Tagliamonte & Baayen, 2012: 164). For example, it might be that DoM is a very useful predictor within a particular NO or disp range but not with others, or that NVar may help discriminate between interesting and uninteresting compounds with adjectival modifiers but not with nominal modifiers. Such interactions may be captured and visualised in a cit.

One should, however, be wary of the fact that cits may also camouflage the contribution of important predictors. The effect of one important variable may for instance be overshadowed by another. A cit-analysis will therefore benefit from a supplementary *random forest analysis*, which is also described in detail in Supplementary Material Online. A random forest analysis computes a conditional importance score for every predictor based on their association with the response variable. Hence, a random forest analysis will detect important variables that may be camouflaged in a cit-analysis.[11]

In the following cit analyses, the response is configured as logarithm to base 10 of look-up regularity + 1 (henceforth *logged look-up regularity*).

In the following, I will perform cit- and random forest analyses to single out the variables and variable interactions that should be included in a look-up based alternative model for compound selection. The results of the analyses will be tested on an independent set of data in Section 6.

5.2. Deriving a new model

5.2.1 Non- and semi-nominal compounds

Since much attention has been devoted to nominal compounds in previous studies (see e.g. Schäfer 2018), I will analyse purely nominal and semi- or non-nominal compounds separately. To begin with the latter, the tree in Figure 8 is generated with a minimum split size of 40, and a significance threshold (henceforth *alpha*) of 0.05 (cf. Supplementary Material Online).

As shown in Figure 8, the uppermost split is based on dispersion, where compounds with a disp > 0.468 are grouped on the right, and the rest are grouped on the left. We can inspect the boxplots below each of these splits (namely Nodes 3, 5 and 6 on the left and Nodes 9, 10

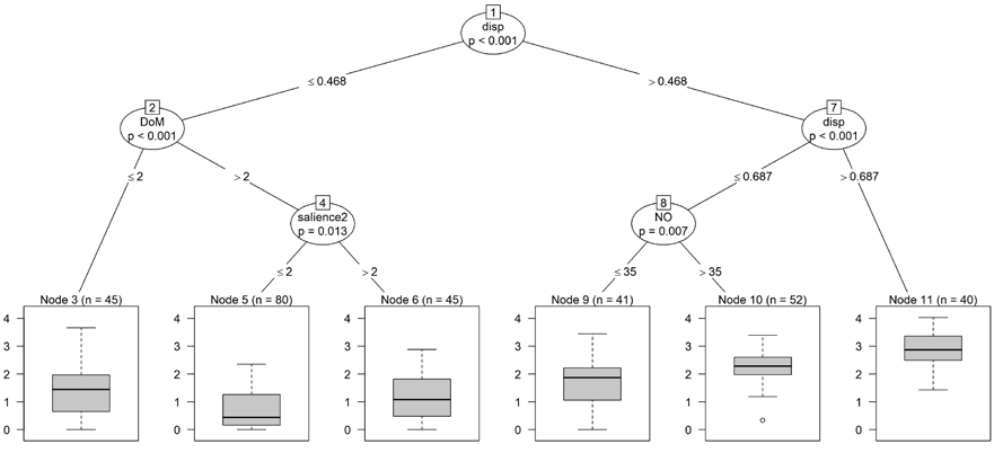


Figure 8: Conditional inference tree predicting the logged look-up regularity of non- and semi-nominal compounds at different levels of the predictors (n = 303).

and 11 on the right) and verify that the ones on the right contain compounds that on average have a higher look-up regularity. In fact, the compounds on the right all have a logged look-up regularity median close to or higher than 2, which corresponds to 99 when we reverse the logarithm. This indicates that dispersion > 0.486 among non- or semi-nominal compounds is associated with medium to high user interest. Since it is not the aim of the new model to differentiate between compounds of various high interest, dispersion > 0.468 is adopted as a qualifying criterion in the new model. See further detailed cit-analysis of non- and semi-nominal compounds with dispersion ≤ 0.468 in Supplementary Material Online. The findings from the multiple cit-analyses over non- and semi-nominal compounds suggest that we may formulate a model based on dispersion, DoM, salience and part of speech that can extract a vast majority of the non- and semi-nominal groupings that at least on a group level appear to be interesting from a user perspective. This will be tested in Section 6.

5.2.2 Nominal compounds

Nominal compounds constitute approximately 75% of the compound data. In the following, I will inspect this section of the data using cits.

Figure 9 contains a cit with a minimum split size of 60 and alpha = 0.05. The cit is complex with many splits. However, if we start by inspecting the boxplots, only two of these, Nodes 3 and 13, present groupings of compounds where the box does not include 0. The groupings in these nodes seem to contain fewer compounds that are not looked up at all than the others, and they have a range that approaches logged look-up regularity of 3 (which corresponds to 999). This latter point is also true for Node 12 and 5. I will nevertheless select the groupings in Node 3 and 13 as compounds that the new model should include, based on the fact that Node 13 has the highest median look-up regularity, and that Node 3 invokes DoM which we already saw was important for the non- and semi-nominal compounds. We may therefore remove compounds with **NO > 61 AND dispersion > 0.613**, and **NO ≤ 61 AND DoM ≤ 2** in order to inspect the remaining compounds further. This inspection involves a random forest analysis and a contingency table and is reported in Supplementary Material Online. What these analyses show is that regardless of which level one chooses, there is no combination of variables that exhaustively predicts the look-up regularity of nominal compounds. However, if one wishes to tune the variables at hand in the new model, the only way to exclude seemingly uninteresting compounds is to also exclude interesting ones, i.e., increasing the specificity of the model comes at the cost of sensitivity, and vice versa.

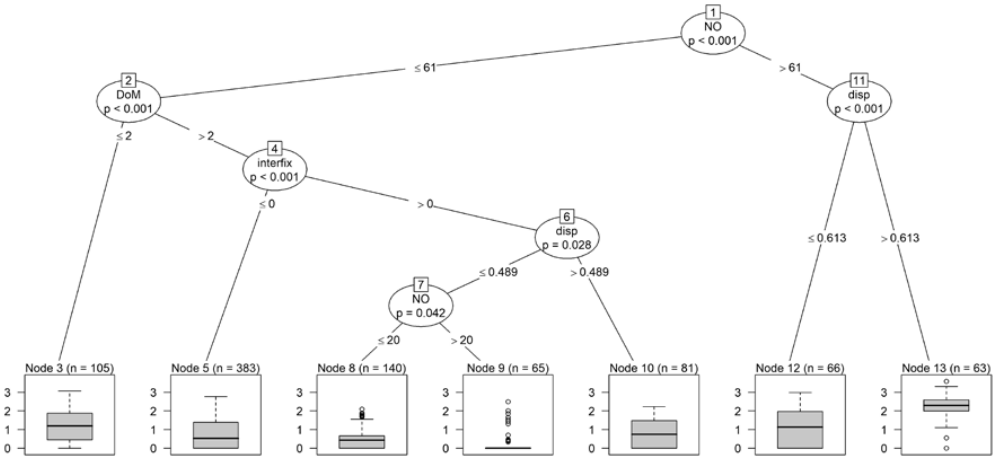


Figure 9: Conditional inference tree predicting logged look-up regularity of nominal compounds at levels of independent variables (n = 903).

The analyses conducted in Sections 5.2.1 and 5.2.2 (and detailed in Supplementary Material Online) single out variables and levels thereof that optimise the proportion of 'interesting' or 'regularly looked up' compounds accepted by the new model. This model is stated and tested in the following.

5.3. The Look-Up Predictor Model

The analyses in Section 5.2. suggest a set of very precise discriminatory levels for the most informative variables. These are accepted without rounding in LookMod that is stated in full below:

Include non- and semi-nominals with the following specifications:

- disp > 0.468
- DoM ≤ 1
- POS-h = adv, n, prep or v AND
  - Salience > 2

OR

  - POS-m = adv, n or preposition

Include nominals with the following specifications:

- disp > 0.4
- NO > 42 AND disp > 0.507
- NO > 100 AND disp > 0.2
- NO ≤ 61 AND DoM ≤ 2
- disp > 0.2 AND interfix = 0

This model will now be tested on both the data from which it was formulated and on an independent set of compound data. Its performance will be evaluated using lacuna analysis in the same manner that StandMod was evaluated in Section 4.2.

6. Testing the Look-Up Predictor Model

In order to test LookMod, we construct an algorithm that runs through the compounds in the data and assigns the value 1 to those which fulfil one or more of its criteria and the value 0 to those which do not.

Table 2 shows the amount of included and excluded compounds in both models at different levels of look-up regularity. Using the number of unvisited entries and lacunas as performance indicators, we can now compare the performance of the two models.

LookMod produces a slightly higher number of unvisited entries (56%) than StandMod (47%). Further, the numbers for 1–10 are very similar across the two models, while StandMod shows better performance in the 10–100-range. Lastly, LookMod includes a few more of the compounds with look-up regularity > 100.

If we define desirables as compounds with a sufficient user-interest to be included in the dictionary, then the lacuna rate would be the rate of unlisted lemmas among the desirables. For the purpose of this study, I will operationalise desirables as compounds with look-up regularity > 10. This means that LookMod has 108 unlisted desirables, which gives a lacuna rate of 18.3%. In comparison, the StandMod has only a slightly better coverage with a lacuna rate of 16.8%. The performance of the StandMod is thus only a little bit more accurate than LookMod on this particular dataset, when it comes to both the number of unvisited entries and the lacuna rate. In other words, StandMod has a slightly higher sensitivity and specificity. One would however expect the StandMod to have a certain advantage since it is the product of many rounds of revision with a case-by-case consideration of each compound, while the LookMod is automatically generated using a handful of absolute criteria. On the other hand, the LookMod is based on look-up regularity which in the current procedure is also employed as the evaluation variable. On this background, while LookMod shows promising results in performing at nearly the same level as StandMod, we should perhaps require an even more convincing advantage in performance if LookMod is to become the new standard.

But there is still a potential source of error that we should control for. LookMod has so far been tested on the very same data from which it originated, which means that it might be overfitted to this particular data. To control for this, we should test the models on a different dataset and compare their performances on that data. This will be done in the following.

6.1. Testing the models on an independent dataset

The independent dataset (henceforth *testset*) is harvested using the same criteria as the initial dataset in Section 3.1 and consists of 214 compounds from the segment *gjerdesitting – glasur*, which along with the other segments in this study has undergone a recent revision.[12] Approximately 77% of testset is currently listed in the BOB. A lacuna analysis will now be performed to assess the performance of both StandMod and LookMod on the testset.

The performance of StandMod and LookMod with respect to look-up regularity in the testset is summarised in Table 3. The proportions of listed lemmas for each level of look-up regularity resemble the proportions of listed lemmas among the original dataset for both models. The inclusion rates are distinctly higher in the desirable levels of look-up regularity for both models, although also a majority of the undesirable compounds are listed by both models. On basis of the testset, LookMod produces a slightly lower proportion of unvisited entries (62%) than StandMod (70%), which contrasts with the results from the original dataset. The difference is however too miniscule (4 compounds) to reflect any true difference between the models. When it comes to the other indicator of performance, namely lacunas, there are 79 desirables in the testset. StandMod captures 72 (91%) of these. In other words, it has 7 lacunas (9%), which is a lower lacuna-rate than for the original dataset. In comparison, LookMod captures 67 (85%) of the desirables, leaving 12 lacunas and a lacuna-rate of 15%. This is also lower than on the original data, but it still

Table 3: Performance of both models on testset

look-up regularity		0		1-10		10-100		>100		
		count	prop %	count	prop %	count	prop %	count	prop %	Sum
StandMod	0	14	30	27	31	6	12	1	3	48
	1	33	70	61	69	45	88	27	97	166
LookMod	0	18	38	39	44	10	20	2	7	69
	1	29	62	49	56	41	80	26	93	145

tells us that StandMod, if anything, performs slightly better than LookMod with respect to lacunas, and thereby sensitivity, for both datasets. Although it is somehow expected that StandMod should perform relatively well as it has come about through meticulous choosing and picking, and it therefore is a promising feature of LookMod that it performs at nearly the same level of specificity and sensitivity, we should demand better accuracy for LookMod if it is to become a lexicographic standard.

6.2. Summary of the performance of the models

It can be inferred from the results in the previous paragraphs that LookMod seems to be a viable starting point for collecting compounds. It picks out a vast majority of the compounds that can be viewed as desirables from a look-up perspective, while keeping the number of unvisited entries at a reasonably low level, at least if we compare it to StandMod. Although StandMod seems to be slightly more accurate with respect to both unvisited entries and lacunas, LookMod is advantageous due to the simple fact that its variables and configurations are explicitly formulated. This is not to say that we are completely in the dark with respect to the inner workings of StandMod, but it does not consist of an explicit set of criteria that can be mechanically applied to a set of compounds - which at least some of the variables in LookMod can.

Furthermore, one must expect a certain "dictionary effect" in the datasets that are utilised in this study. Since the dictionary in question is undergoing a revision, it is likely that a certain portion of the look-ups are conducted by the lexicographers working with the revision. This might cause a slight inflation of the look-up frequency of the listed lemmas, which would have a positive effect on the performance of StandMod as it is evaluated here.[13] It should however be noted that this effect is somewhat controlled for by the dispersion estimate that is incorporated in the look-up regularity variable. LookMod is affected by the dictionary effect via the look-up regularity scores that its criteria are derived from.

All in all, the primary utility for LookMod is probably not to be employed as an autonomous machine that creates lemmalists unaided by humans. And its accuracy rate does not at this point warrant a complete refurnishing of lexicographic practice. Rather, its usefulness comes as a working tool for lexicographers and lexicologists. The procedure that has been undertaken in order to generate the LookMod provides valuable information about the usefulness of the variables in question. It also indicates that some variables are not universally useful but have applicability within subsets of compounds. Among these are degree of motivation, which is an important predictor among non- and semi-nominal compounds and infrequent nominal compounds, salience which has applicability within certain parts of speech after controlling for dispersion, and number of occurrences whose primary utility is among compounds in the 0.2-0.4 dispersion range. Dispersion then, is the only variable in this study that seems to be a globally important predictor of look-regularity.

7. Concluding remarks

In this study, several methods have been applied to a material of 1206 Norwegian compounds. The aim has been to evaluate the current lemmalist of compounds in *Bokmålsordboka* and devise an alternative model for lexicographic compound selection. The Standard and Look-Up Predictor models have been evaluated using look-up statistics from the two online dictionaries *Bokmålsordboka* and *Nynorskordboka*.

One might argue that it makes little sense to operate with a set of variables that are only able to a certain extent to predict the look-up distribution of a certain compound, when one can easily consult the look-up counts directly. However, not all lexicographical or lexicological projects can benefit from an existing body of look-up information. Additionally, and perhaps more importantly, the look-up statistics only show what has been looked up in the past, and not what one can expect to be looked up in the future. Although one might expect the user interest of yesterday to resemble the user interest of tomorrow, a set of variables might be able to cover both the commonalities and the disparities between the two, whereas strict adherence to the direct look-up statistics can only cover the former.

Of course, the ideal is not to create an autonomous machine that selects compounds unaided by humans, but rather to supply lexicographers with working tools. No doubt, one can only expect that LookMod would be even more precise, and possibly outperform StandMod, if it were supplemented with the critical assessment of a group of lexicographers. Furthermore, the utility of the LookMod is not necessarily as a meticulous procedure that must be complied with to the letter, but as a sensible list of variables and levels thereof. It also conveys information about the hierarchy of variables, for instance that dispersion is a globally important variability, while DoM's and NO's utilities as variables are chiefly among slightly underdispersed compounds.

There is also something to be said for not sticking too closely to the arbitrary variations and idiosyncrasies of look-up statistics, but rather to conform to rigorous selectional variables that reflect broader patterns of look-up behaviour. Look-ups can for example be catalysed by temporarily socially relevant things such as crossword puzzles, the news cycle, seasons and public holidays (see Bäckerud, Nilsson & Sköldberg (2020) and Wolfer et al. (2014)). The BOB for instance is accessed a lot in connection with nationwide high school exams, where a given compound in the handout materials at such an exam may catalyse thousands of look-ups for that compound. Such arbitrary effects call for a moderate use of look-up frequency as an indicator of word importance and stress the importance of look-up dispersion over time.

Data on look-up behaviour in online dictionaries has a wide range of research possibilities. With respect to the question of compound selection, and especially the current unexplained variation among nominal compounds, future research could for example include a wider range of qualitative variables or distributional measurements from more than one corpus. Such adjustments might contribute to make more accurate predictions of look-up behaviour.

Finally, meta-information about the users performing the look-ups might help to uncover what needs different user groups have, and how one can design the lemmalist to meet those needs.

Notes

1 The search logs that will be inspected for this study are drawn from the website *ordbok.uib.no* which has an interface that enables users to make parallel queries in the official Norwegian dictionaries for the two written standards of Norwegian, *Bokmål* and *Nynorsk*. The interface may therefore be utilised as a bilingual dictionary to check equivalency between the two standards.

2 Müller-Spitzer et al. (2015) also finds that polysemic words are more frequently looked up than monosemic ones, even when controlling for the fact that the most frequent words are polysemic. I will not

include polysemy as a factor here, because the data consist of many words that are not listed in dictionaries. Therefore, I have no objective way to determine what is mono- or polysemic.

3 This means that derivations of compounds such as *tospråklighet* "bilingualism" (lit. "twolingualness") do not count as compounds since there is no way to divide it into two individual stems because neither *språklighet* lit. "lingualness" or *-het* "-ness" are stems on their own.

4 Leaving affixes, symbols, abbreviations and the like aside.

5 The dictionaries were formerly accesible through an interface located at *ordbok.uib.no* and the look-up stats in question are gathered from this old interface.

6 It should be noted that the regular expression has some caveats, for instance that people may use URLs as query expressions and thereby obscure the textual surroundings that normally enclose the search query. These shortcomings are however not expected to have any influence on the results of this study.

7 https://ordbok.uib.no/stats/h/mest.sokt.2.html

8 The reason why not every query is included is mainly to filter out noise and hapax legomena that only serve to slow down the computational processes involved in obtaining the look-up statistics. Besides, a handful of missed queries here and there does not alter the general tendencies of the look-up frequency variable.

9 These two measures have opposite scales, but I have reversed the scale of DP so that it aligns with Juilland's D and Number of occurrences.

10 More specifically, I performed a stepwise regression procedure of a generalised linear binomial model where all variables in the study, including interactions between NO, disp and the qualitative variables, were included. The procedure suggested, based on the Aikake information criterion, a model with disp as its sole predictor of BOB status. Fitting such a model showed that an increase in disp is associated with a higher likelihood of a compound being included in the dictionary. This information can also be easily obtained from Figure 5.

11 The conditional inference tree and random forest procedure are performed by using the Cforest function of the party package in R (Hothorn et al. 2006, R Core Team 2017).

12 The lexicographer that has edited the testset has also edited one of the segments in the original dataset.

13 It is of course hard to pinpoint exactly how often editors visit the dictionary home page and hence how large the dictionary effect is, but from my own experience of editing BOB, the dictionary home page is visited several times a day. It is not unlikely that a single editor is responsible for 10+ look-ups of the same compound lemma.

References

A. Dictionaries

Bokmålsordboka. *Språkrådet and University of Bergen*. (ordbokene.no).
Norwegian Academy Dictionary. *Det Norske Akademi for språk og litteratur*. (naob.no).
Nynorskordboka. *Språkrådet and University of Bergen*. (ordbokene.no).

B. Other literature

Bäckerud, E., Nilsson, P. and E. Sköldberg. 2020. 'Så används Svenska Akademiens ordböcker på nätet. Implicit och explicit feedback från användarna.' *Nordiske Studier i Leksikografi* 15: 91–101.

Fjeld, R. V., Nøklestad, A. and K. Hagen. 2020. *'Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk.'* In Leksikografi og korpus. En hyllest til Ruth Vatvedt Fjeld. Edited by Johannessen, J.B. and K. Hagen. *Oslo Studies and Language* 11: 47–59.

Fjeld, R. V. and L. Vikør. 2008. *'Ord og ordbøker.'* Høyskoleforlaget.

Gries, S. T. 2008. 'Dispersions and Adjusted Frequencies in Corpora.' *International Journal of Corpus Linguistics* 13: 403–437.

Hothorn, T., Hornik, K. and A. Zeleis. 2006. 'Unbiased Recursive Partitioning: A Conditional Inference Framework.' *Journal of Computational and Graphical Statistics* 15: 651–674.

Juilland, A. G., Brodin, D. R. and C. Davidovitch. 1971. *'Frequency Dictionary of French Words.'* Mouton de Gruyter.

Kulbrandstad, L. and T. Kinn. 2016. 'Språkets mønstre'. 4th edition. Universitetsforlaget.

Leksikografisk bokmålskorpus. Distributed by the CLARINO. UiB portal: https://clarino.uib.no/lex/corpus/concordance

Levshina, N. 2015. *'How to Do Linguistics with R: Data Exploration and Statistical analysis'*. John Benjamins Publishing Company.

Müller-Spitzer, C., Wolfer, S. and A. Koplenig. 2015. 'Observing Online Dictionary Users: Studies Using Wiktionary Log Files'. *International Journal of Lexicography* 28: 1–2.

Paulsen, M. E. 2022. 'Assessing Word Commonness: Adding Dispersion to Frequency'. *International Journal of Corpus Linguistics*. (https://doi.org/10.1075/ijcl.21037.eke).

R Core Team. 2017. *'R: A Language and Environment for Statistical Computing.'* R Foundation for Statistical Computing. Vienna, Austria.

Schryver, G.-M., Joffe, D., Joffe, P. and S. Hillewaert. 2006. 'Do Dictionary Users Really Look Up frequent words? On the Overestimation of the Value of Corpus-Based Lexicography.' *Lexikos* 16: 67–83.

Schryver, G.-M. d., Wolfer, S. and R. Lew. 2019. 'The Relationship between Dictionary Look-Up Frequency and Corpus Frequency Revisited: A Log-File Analysis of a Decade of User Interaction with a Swahili-English Dictionary.' *Gema Online Journal of Language Studies* 19: 1–27.

Schäfer, M. 2018. *'The Semantic Transparency of English Compound Nouns.'* Language Science Press.

Strobl, C., Malley, J. and G. Tutz. 2009. 'An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests.' *Psychological Methods* 14.4: 323–348.

Svanlund, Jan. 2002. 'Lexikalisering.' *Språk och stil* 12. 7–45.

Tagliamonte, S. A. and R. H. Baayen. 2012. 'Models, Forests, and Trees of York English: Was/were Variation as a Case Study for Statistical Practice'. *Language Variation and Change* 24: 135–178.

Trap-Jensen, L., Lorentzen, H. and N.H. Sørensen. 2014. 'An Odd Couple – Corpus Frequency and Look-Up Frequency: What Relationship?' Slovenščina 2.0, 2.2. Edited by: Kosem, I and M. Rundell. Trojina, Institute for Applied Slovene, Slovenia: 94–113.

Wolfer, S., Koplenig, A., Meyer, P. and C. Müller-Spitzer. 2014. 'Dictionary Users Do Look Up Frequent and Socially Relevant Words. Two Log File Analyses.' Proceedings of the 16th EURALEX International Congress. Edited by Abel, A., Vettori, C., and N. Ralli. Eurac research, Bolzano: 281–290.

Grafisk design: Kommunikasjonsavdelingen, UiB / Trykk: Skipnes Kommunikasjon AS

uib.no

9788230861998 (print)
9788230855836 (PDF)