Hypotesetesting versus utforskende statistikk i språkforskning

Bård Uri Jensen
Høgskolen i Innlandet

Sammendrag

Denne artikkelen handler om forholdet mellom hypotesetestende og utforskende statistiske tilnærminger i språkforskningen, med vekt på andre- og flerspråklighetsforskning. Den bruker eksempler fra publiserte studier til å peke på noen utfordringer ved hypotesetesting, som kravet til spesifikke og eksplisitte hypoteser, premisset om uavhengige observasjoner og prinsippet om gruppevis feilrate, og viser hvordan noen slike utfordringer kan møtes ved å endre tilnærmingen til analysene. Videre argumenteres det for at fokuset på signifikans og *p*-verdier kan ha blitt for sterkt i deler av språkforskningen, og at vi vil være tjent med å rapportere resultater på andre måter og supplere med andre analysetilnærminger. Artikkelen demonstrerer et knippe av enkle visualiseringsteknikker og trekker frem deres verdi som verktøy i en initiell analysefase, og forklarer også på et overordnet nivå tre mer avanserte multivariate utforskende analysemetoder og deres potensielle rolle som hypotesedannende verktøy. Artikkelens hovedpoeng er hvordan hypotesetestende og utforskende tilnærminger inngår i et samspill, der begge spiller nødvendige roller i språkforskningen.

Nøkkelord: *kvantitative analyser*; *hypotesetester*; *utforskende tilnærminger*; *signifikans*; *visualisering*; *multivariate metoder*

Innledning

Denne artikkelen diskuterer bruken av ulike statistiske metoder i språkforskning. Den vektlegger samspillet mellom hypotesetestende og utforskende tilnærminger og benytter eksempler fra ulike områder av språkvitenskapelig forskning. Artikkelen er bygd opp på følgende måte: Etter en liten, illustrerende fortelling gis en kort innføring i henholdsvis hypotesetesting og utforskende tilnærminger. I det følgende kapitlet blir det vist hvordan disse tilnærmingsmåtene kan inngå i et samspill. Deretter drøftes tre eksempelstudier der hypotesetesting er brukt på uheldige måter, og hvordan man i hvert tilfelle kunne ha tilnærmet seg disse problemstillingene på en bedre måte. Neste kapittel viser først noen enkle visualiseringsteknikker som kan brukes i initiell dataanalyse, før et par multivariate analysemetoder presenteres. Til slutt kommer en kort oppsummering og diskusjon. Artikkelen er en bearbeiding av min prøveforelesningen over oppgitt tema til ph.d.-graden ved Universitetet i Bergen 4. desember 2017.

En liten, fiktiv forhistorie

Da jeg gikk over Torgallmenningen i Bergen før prøveforelesningen min, la jeg merke til noe rart. Jeg hørte masse skarre-r-er rundt meg. Og jeg tenkte umiddelbart:

(1) Er det slik at skarre-r er mer vanlig enn rulle-r i Bergen?

Dette krevde åpenbart en vitenskapelig undersøkelse. Jeg formet en hypotese om at skarre-r er mer vanlig i Bergen enn rulle-r, eller at det i populasjonen av mennesker som bor i Bergen, er flere språkbrukere som skarrer enn som ruller på r-ene. For å kunne teste denne hypotesen konstruerte jeg en null-hypotese, nemlig at skarre-r og rulle-r er like vanlig, eller at det ikke er noen forskjell i andel språkbrukere som bruker skarre-r, og andel språkbrukere som bruker rulle-r. Så gjorde jeg et lite utvalg. Jeg satte meg på en benk og merket meg hvordan de 5 neste personene som gikk forbi, realiserte r-ene sine. Resultatet var at av de 5 personene i utvalget var det 5 som brukte skarre-r og ingen som brukte rulle-r.

For å analysere dette resultatet tar man utgangspunkt i en antagelse om at nullhypotesen er sann. I dette tilfellet vil det si at de to gruppene av språkbrukere utgjør en like stor andel av språkbrukerne i Bergen. Hvis de er like vanlige, så er sjansen for at en tilfeldig utvalgt person skarrer, lik 50 %, eller 0.5. Altså var sannsynligheten for at den første personen i utvalget mitt skulle skarre, lik 0.5. Sannsynligheten for at den neste personen i utvalget skulle skarre, var også 0.5.[1] Sannsynligheten for at begge disse skulle skarre, kan man regne ut ved å multiplisere de to sannsynlighetene med hverandre, altså 0.5 multiplisert med 0.5, lik 0.25. Slik kunne man fortsette, og man kan regne ut sannsynligheten for at *alle* de 5 personene i utvalget var skarrere, ved å multiplisere 0.5 med seg selv 5 ganger. En slik sannsynlighetsverdi kalles gjerne for p (for *probability*), og regnestykket og svaret vises i (2).

(2) $p = 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.03125$

Ifølge vanlig praksis i humanistisk forskning ville man bruke denne lave sannsynligheten, eller p-verdien, til å forkaste nullhypotesen. Resultatet ville dermed være en støtte til hypotesen om at en større andel av språkbrukerne i Bergen skarrer enn ruller på r-ene.

Hypotesetesting

Det lille, fiktive eksemplet ovenfor illustrerer de generelle prinsippene for statistisk hypotesetesting i språkforskning og for så vidt også i hvilken som helst annen forskning. Prosedyren følger de samme stegene i denne rekkefølgen. Formålet er å bekrefte eller avkrefte en hypotese man har om en *parameter* i en *populasjon*, altså at det finnes en form for tendens i populasjonen. I språkforskning er populasjonen gjerne en befolkning av språkbrukere, som for eksempel alle mennesker som bor i Bergen, men den kan bestå av andre typer enheter. Tendensen som man har en hypotese om, kan for eksempel være at det er en systematisk forskjell i verdiene mellom to grupper i populasjonen, eller at det finnes en sammenheng blant individene i populasjonen mellom to variabler. Man

[1] Egentlig er denne sannsynligheten ørlite lavere enn 0.5, ettersom én skarrer allerede er plukket ut fra populasjonen. I store populasjoner spiller denne forskjellen så liten rolle at den kan ses bort fra.

konstruerer en nullhypotese, som går ut på at tendensen i hypotesen ikke finnes, altså at parameteren har verdien 0. Man gjør et tilfeldig utvalg blant individene i populasjonen og måler styrken i tendensen i utvalget, altså et *estimat* på den reelle parameteren. Her bryter eksempelundersøkelsen av skarring med forutsetningene, ettersom utvalget i den undersøkelsen ikke var tilfeldig og neppe kan hevdes å være representativt for populasjonen, i og med at bare mennesker som gikk over Torgallmenningen en bestemt ettermiddag, ville kunne bli med i utvalget. Man regner ut sannsynligheten for å finne en tendens i utvalget som er minst så sterk som den man har funnet, under antagelsen at nullhypotesen er sann. Man sammenligner den sannsynligheten med et kritisk nivå som man har valgt seg på forhånd, ofte kalt signifikansnivået eller α-nivået. Dersom den sannsynlighetsverdien man har regnet ut, er lavere enn det valgte kritiske nivået, *forkaster* man nullhypotesen; man regner den som tilbakevist eller usannsynliggjort. I så fall regner man resultatet som en støtte til eller en bekreftelse av den egentlige hypotesen, og man kaller resultatet *signifikant* og rapporterer den aktuelle tendensen som en egenskap ved populasjonen. Et vanlig α-nivå i humanistisk forskning er 0.05, eller 5 %, og dette gjenspeiler da det nivå av usikkerhet som vi synes er akseptabelt når vi rapporterer om et positivt funn. Det er vesentlig å være oppmerksom på denne statistikkfaglige betydningen av ordet 'signifikant'; det betyr utelukkende at resultatet er sannsynlig innenfor et predefinert nivå av sikkerhet, og sier ingenting om hvorvidt resultatet er betydningsfullt eller viktig.

Dette er den grunnleggende logikken og prosedyren for all hypotesetesting, der man følger disse stegene i denne rekkefølgen og regner ut en *p*-verdi ved hjelp av en matematisk formel. De hypotesetestene vi gjerne kjenner under navn som for eksempel Students *t*-test eller Pearsons korrelasjonstest, er rett og slett matematiske formler for å regne ut *p*-verdier som så kan sammenlignes med α. Forskjellen er at formlene i disse testene er noe mer komplekse enn formelen vi kunne bruke i skarre-eksemplet, og man benytter derfor gjerne et dataprogram til å gjøre beregningene.

Det er viktig å være bevisst at hypotesetesting betyr testing av hypoteser. Dette kan virke som en tautologi eller en banalitet, men det er mitt inntrykk at det blant språkvitere ikke er så uvanlig å overse dette, og det kan ha uheldige konsekvenser for forskningens gyldighet. Når man bruker en hypotesetest, skal den alltid ta utgangspunkt i en konkret,

formulert hypotese som skal testes. Hypotesen må være formulert på forhånd, altså før datainnsamlingen, eller i det minste før dataanalysen er påbegynt. Hvis man ser på dataene først og formulerer hypoteser etterpå, blir logikken i beregningen av *p*-verdien forrykket, og resultatene fra eksperimentet eller undersøkelsen vil være ugyldige. Man bruker altså hypotesetester til å bekrefte eller avkrefte hypoteser, og tilnærmingen blir gjerne også kalt for bekreftende (*confirmatory*) dataanalyse.

Utforskende tilnærminger

Ikke alle kvantitative forskningsstudier tar utgangspunkt i formulerte hypoteser. Noen ganger undersøkes variabler eller studieobjekter som det finnes lite eller ikke noe grunnlag for å formulere hypoteser om. Da kan man bruke statistiske teknikker for å *utforske* dataene. *Utforskende* eller eksplorative statistiske tilnærminger dreier seg om dataanalyser som med ulike metoder og verktøy har som formål å utforske, gjøre seg kjent med, få overblikk over data, altså egenskapene til variabelfordelinger eller kombinasjoner av variabelfordelinger.

Denne typen økt kunnskap og forståelse om data kan være et mål i seg selv, eller den økte kunnskapen kan i neste trinn danne et grunnlag for å utvikle nye hypoteser som deretter kan testes på nye data. Utforskende tilnærminger brukes dessuten gjerne i en tidlig fase i en undersøkelse for å utvikle et kunnskapsgrunnlag for å velge gyldige hypotesetestende analysemetoder. De statistiske teknikkene som benyttes for å utforske data, som visualiseringer eller multivariate metoder, er typisk andre enn dem man bruker til hypotesetesting; dette kommer jeg tilbake til.

Et eksempel på en utforskende tilnærming er fra Gujord, Halverson og Jensen (2017), som studerer 8 variabler (se tabell 1 neste side) knyttet til realisering av subjekt i tre typer tekster på norsk.

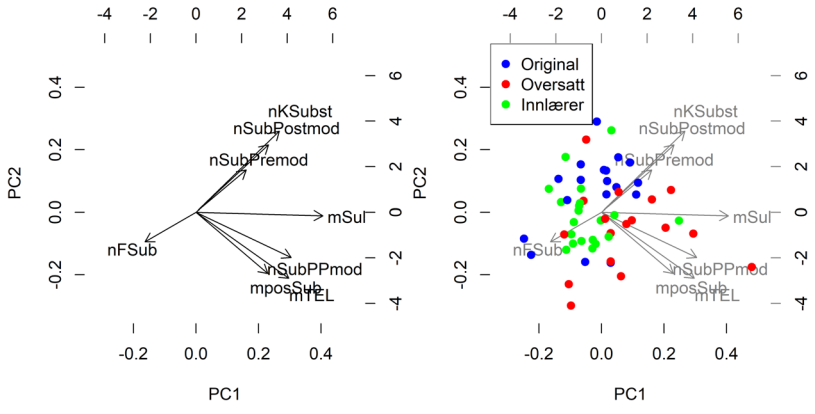
De tre teksttypene er norske originaltekster, tekster oversatt fra engelsk, og tekster skrevet av innlærere med engelsk som førstespråk, totalt et utvalg på 55 tekster. Gujord og kollegene hennes analyserer så de 8 variablene med en multivariat metode kalt prinsipalkomponentanalyse (PCA). Dette er en analysemetode som man bruker til å redusere kompleksiteten i et sammensatt rom av mange variabler, slik at så mye som mulig av variasjonen i materialet samles i få dimensjoner. I dette tilfellet

Tabell 1: Egenskaper ved subjektet, brukt i prinsipalkomponentanalyse (Gujord et al. 2017).

ID	Variabel
mTEL	Gjennomsnittlig t-enhetslengde (antall ord)
mposSub	Gjennomsnittlig subjektsposisjon i t-enheten (antall ord fra starten)
nFSub	Andel av subjektene som er formelle subjekter (ikke-referensiell 'det')
mSubL	Gjennomsnittlig subjektslengde (antall ord)
nKSubst	Andel av subjekter som har substantiv i kjernen
nSubPremod	Andel av subjekter med premodifisert kjerne
nSubPostmod	Andel av subjekter med postmodifisert kjerne
nSubPPmod	Andel av subjekter med både pre- og postmodifisert kjerne

tar analysen utgangspunkt i et rom av 8 variabler som er dannet av 55 observasjoner på hver av disse variablene. Rent teknisk går prinsipalkomponentanalysen frem på den måten at den først samler så mye variasjon eller informasjon som mulig i den første dimensjonen ved å optimalisere en vektet lineær kombinasjon av alle de ulike variablene. Deretter samler den så mye som mulig av den gjenværende variasjonen i en ny dimensjon som er uavhengig av den første, og som altså ikke korrelerer med den. Slik går den frem inntil den har gått gjennom alle de 8 dimensjonene (Everitt og Hothorn 2011:61–62). I de siste dimensjonene vil det dermed gjenstå svært lite variasjon, og materialet kan beskrives nesten like godt uten å ta hensyn til disse dimensjonene. Hvor godt man kan beskrive materialet med få dimensjoner, avhenger av hvordan variasjonen i materialet arter seg, men i praksis vil ofte bare to eller tre dimensjoner benyttes for å gi en oversikt, selv om oversikten da ikke gjengir det fullstendige bildet. Gevinsten ligger i en enklere beskrivelse av et komplekst materiale i et lite antall nye variabler (altså dimensjonene), som dessuten er uavhengige av hverandre.

I det venstre diagrammet i figur 1 vises de to første dimensjonene fra prinsipalkomponentanalysen til Gujord et al. Disse to dimensjonene representerer i dette tilfellet til sammen noe over halvparten av den samlede variasjonen i materialet. I et slikt diagram fra en prinsipalkomponentanalyse representerer lengden på pilene hvor sterkt de ulike variablene bidrar, og retningen viser deres bidrag til de to dimensjonene og hvordan variablene ligner hverandre og skiller seg fra hverandre i disse to dimensjonene.



Figur 1: En grafisk fremstilling av de to første dimensjonene fra en prinsipalkomponentanalyse av datamaterialet til Gujord et al. (2017). Til venstre vises hver variabels bidrag til de to dimensjonene. Til høyre de samme to dimensjonene, der blå prikker representerer originaltekstene, røde prikker de oversatte tekstene, og grønne prikker innlærertekstene. Se tabell 1 for forklaring av de enkelte variablene.

Når det todimensjonale rommet er dannet, kan de opprinnelige observasjonsenhetene plasseres på de to aksene, og man kan se hvordan enhetene ligger i forhold til hverandre i dette todimensjonale rommet. I det høyre diagrammet i figur 1 er hver tekst representert med et punkt i bildet, der hver teksttype har fått sin egen farge. Prinsipalkomponentanalysen kjenner altså ikke til disse tre teksttypene, så analysen utnytter ikke informasjonen om at det er tre ulike teksttyper i materialet; den kjenner bare til verdien av de 8 språklige variablene for hver av de 55 tekstene. Men diagrammets forenklede fremstilling av det komplekse materialet hjelper forskerne til å se at de tre teksttypene ganske tydelig plasserer seg i tre ulike regioner i det todimensjonale feltet som analysen danner. Tilnærmingen kan altså for det første tilby et umiddelbart inntrykk av *at* de tre teksttypene synes å ha noen typiske egenskaper som skiller dem fra hverandre, og at det også er noen likheter mellom dem. Ved å se nærmere på hvordan de to dimensjonene er definert av de enkelte variablene, kan man for det andre begynne å danne seg et bilde av *hvordan* de tre teksttypene er forskjellige, og hvordan de ligner hverandre. Denne utforskende analysen med tilhørende visualisering

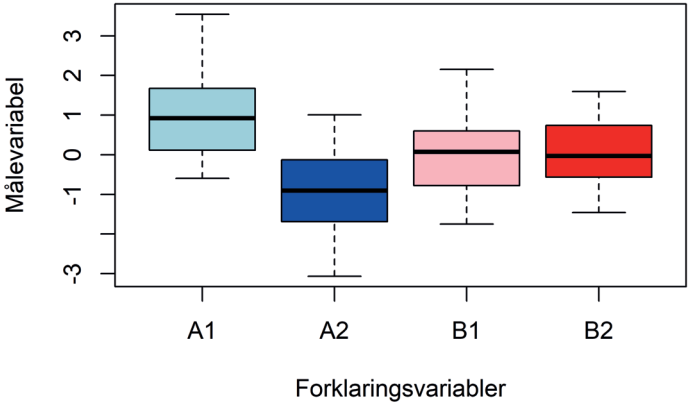
kan dermed bidra til å gi et ganske intuitivt bilde av hvordan disse teksttypene forholder seg til hverandre, uten at det lå noen hypotese om dette til grunn.

I dette konkrete diagrammet i figur 1 ser vi blant annet at originaltekstene har høyere andel subjekter med substantiv i kjernen og subjekter som er utbygd med enten pre- eller postmodifiseringer (*nKSubst*, *nSubPremod*, *nSubPostmod*), mens oversatte tekster blant annet har gjennomsnittlig lengre t-enheter og også gjennomsnittlig plasserer subjektet lengre ut i t-enheten (*mTEL*, *mposSub*), og innlærertekstene har høyere andel av formelle subjekter (*nFSub*).

Samspillet mellom to tilnærminger

Jeg ønsker med denne artikkelen å vise at det ikke er noen motsetning mellom det å bruke hypotesetestende og utforskende tilnærminger til analyse av data, og peke på hvordan de utfyller hverandre og delvis også glir over i hverandre. Jeg vil bruke eksempler fra språkforskning til å belyse de to tilnærmingene, men også til å vise hvordan de er relevante for andrespråkforskning. Mange av eksemplene er hentet fra korpuslingvistikk, men drøftingen er selvfølgelig relevant også for andre typer språkforskning, som for eksempel psykolingvistiske eksperimenter med måling av reaksjonstider eller spørreskjemaundersøkelser om språkholdninger.

Vi skal se på et nytt fiktivt eksempel. La oss tenke oss at det kommer en opprømt språkforsker til meg med diagrammet i figur 2. Forskeren har sittet og studert diagrammer over ulike forklaringsvariabler uten å oppdage noen effekter på målevariabelen, men har nå funnet en kombinasjon av to forklaringsvariabler som gir utslag på målevariabelen i diagrammet. I andrespråksforskningen kunne et slikt diagram for eksempel representere feilfrekvens på en spesifikk språklig variabel (målevariabelen), mens de to forklaringsvariablene kunne være henholdsvis kjønn og førstespråk. I så fall ville A1-gruppen i figuren representere menn med førstespråk 1, mens B2 ville representere kvinner med førstespråk 2, og tilsvarende for A2 og B1. Forskeren ber meg sjekke om den synlige effekten i diagrammet er signifikant, altså om *p*-verdien er lavere enn α:



Figur 2: Boksdiagram som viser interaksjon mellom to dikotome forklaringsvariabler på en fiktiv målevariabel. Den ene forklaringsvariabelen har verdiene A og B, og den andre har verdiene 1 og 2.

(3) Er effekten i figur 2 signifikant?

Det er to problemer med denne anmodningen. Det første problemet, som jeg allerede har vært inne på, er at spørsmålet ikke er dannet med utgangspunkt i en ferdig formulert hypotese. Spørsmålet er dannet etter at dataene er inspisert, og da har man altså ikke noen hypotese, og man følger ikke den nødvendige rekkefølgen i prosedyren for hypotesetesting. Med en hypotese som er basert på observasjoner av de samme data som man tester på, forrykker man grunnlaget for sannsynlighetsberegningen knyttet til det som skal være et tilfeldig utvalg, og resultatet blir ugyldig. I den fiktive observasjonsstudien fra Torgallmenningen fikk jeg derimot først et inntrykk basert på det man kan kalle en liten pilotstudie. Deretter formulerte jeg en hypotese basert på inntrykket fra pilotstudien, og til slutt samlet jeg inn nye data som jeg testet hypotesen på. Slik fulgte Torgallmenning-studien den riktige logikken i forløpet.

Det andre problemet er at spørsmålet i (3) ikke er spesifikt. Diagrammet viser en interaksjon mellom de to forklaringsvariablene, men anmodningen om hypotesetesting presiserer ikke spesifikt hvilken tendens forskeren ønsker å sjekke signifikansen for. Det kan være forskjellen mellom A1 (lyseblå) og A2 (mørkeblå), eller forskjellen mellom A1 (lyseblå) og B1 (rosa), eller forskjellen mellom A1 og resten av

gruppen, eller forskjellen mellom differansen mellom A1 og A2 (de blå) og differansen mellom B1 og B2 (de røde). Alle disse er rimelige hypoteser å ha, og kanskje er de også vitenskapelig interessante og relevante. Men når man skal teste en hypotese, må den være spesifikk. I den fiktive studien av data fra Torgallmenningen ville et tilsvarende, uspesifikt forskningsspørsmål være:

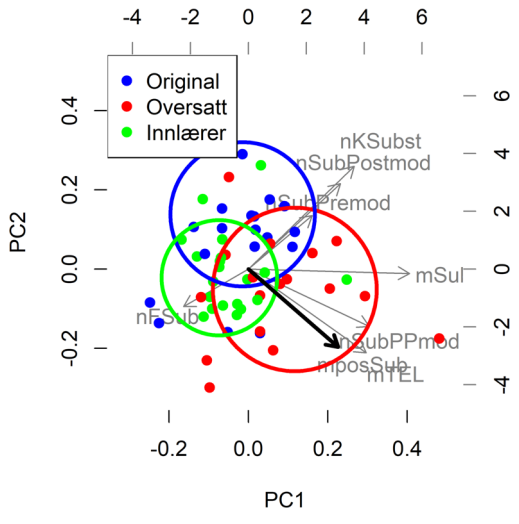
(4) Er det flere som skarrer i Bergen?

Forskningsspørsmålet som jeg faktisk stilte i (1), var om det i Bergen er flere som skarrer enn ruller på r-ene, men en annen tolkning av spørsmålet i (4) ville være om det er en større andel som skarrer i Bergen enn i et annet, spesifisert område av Norge, f.eks. Hamar. Spørsmålet i (3) ovenfor er altså ikke spesifikt nok.

Et sentralt poeng er dermed hvordan man danner hypoteser som er egnet for hypotesetesting. Én tilnærming er å bruke en utforskende analyse på et datamateriale, formulere en hypotese på grunnlag av en tendens man finner i den utforskende analysen, og deretter samle inn et nytt datamateriale som man bruker til å teste hypotesen på. Man kunne for eksempel bruke prinsipalkomponentanalysen i figur 1 og figur 3 til å generere en hypotese om at oversatte tekster på norsk (røde punkter i diagrammene) har subjektet plassert lengre ut i setningen enn originaltekster (blå punkter). Deretter kunne man samle inn et nytt tekstmateriale og teste denne spesifikke hypotesen med en passende hypotesetest, for eksempel Students *t*-test eller tilsvarende. I så fall ville studien basert på prinsipalkomponentanalyse fungere som en pilotstudie for en helt parallell studie med helt tilsvarende studieobjekt og helt tilsvarende variabler.

Vi bruker ofte tidligere empiriske studier som grunnlag for å danne hypoteser. Dette kan være selvstendige studier eller pilotstudier, og de kan være utforskende studier eller hypotesetestende studier. Vi kan også ha et rent teoretisk grunnlag for en hypotese, men det vanligste er kanskje å ha en kombinasjon, eller i det minste at det empiriske grunnlaget også er forankret i en form for teori. Men når vi formulerer hypotesen, må vi sørge for at den er spesifikk og eksplisitt, så vi kan bruke en hypotesetest til å teste den.

Ofte er ikke grunnlaget for å formulere en hypotese like solid som i det tilfellet jeg beskriver ovenfor, men man har kanskje noen empiriske re-



Figur 3: Resultatet av en prinsipalkomponentanalyse (Gujord et al. 2017). Ringene og pila er satt inn manuelt på grunnlag av visuell inspeksjon. Den svarte pila indikerer at de oversatte tekstene (røde punkter) har en plassering i rommet som er lengre ned og til høyre enn de andre tekstene, noe som er i samme retning som variabelen som representerer subjektets posisjon i setningen (*mposSub*).

sultater som dreier seg om et studieobjekt som ikke er helt det samme, eller variabeltyper som ikke svarer helt til de variablene man ønsker å undersøke. I tillegg har man kanskje en teori som knytter de to studieobjektene sammen. Hva som kan sies å være et "godt nok" grunnlag for å danne en hypotese, er et spørsmål uten klare svar. I doktoravhandlingen min undersøker jeg hvordan skriveverktøyet påvirker forskjellige leksikosyntaktiske trekk i elevtekster skrevet på førstespråket (Jensen 2017). Jeg har imidlertid ikke noe empirisk hypotesegrunnlag som er direkte knyttet til denne type data. Derimot har jeg en del empiri fra forskning på registervariasjon, delvis knyttet til teori om at produksjonshastighet er en årsak til forskjeller mellom muntlig og skriftlig språkbruk (Biber 1988; Chafe og Danielewicz 1987; Halliday 1989). Dessuten viser jeg til en gammel studie som handler om hvordan tre ulike ikke-digitale skriveverktøy med ulik hastighet påvirker en enkelt språkvariabel kalt TTR[2] på en slik måte at høyere skrivehastighet

[2] *Type/token ratio*, altså forholdstallet mellom antall typer og antall eksemplarer, i dette tilfellet av ordformer i teksten.

gir TTR-verdier som man gjerne forbinder med mer "muntlig" eller "spontant" språk (Horowitz og Berkowitz 1964). Men jeg har ingen støtte i studier om digitale skriveverktøy som tar hensyn til egenskapene til disse verktøyene. Man kan absolutt sette spørsmålstegn ved et slikt fundament for en hypotese.

Så lenge en hypotese ikke direkte strider mot hovedvekten av den bakgrunnskunnskapen man har, så mener jeg det ikke er noe direkte i vegen for å teste den. Tukey advarer imidlertid sterkt mot å tro at "vitenskap [...] begynner med et ryddig spørsmål" (1980:24), og normalt ønsker vi å kunne argumentere både empirisk og teoretisk for at hypotesen vår representerer en realistisk modell av virkeligheten. Testing av fullstendig grunnløse hypoteser gjør det problematisk både å forklare og drøfte resultatene og eventuelt sette dem inn i en større kunnskapsmessig sammenheng. Dessuten kan testing av grunnløse hypoteser komme i vegen for og redusere styrken i analysene av velbegrunnede hypoteser, slik jeg kommer tilbake til.

Eksempler på studier med hypotesetester

Jeg vil nå se nærmere på studier fra tre doktoravhandlinger som eksempler på hver sin type utfordring knyttet til hypotesetesting, og hver med sin potensielle løsning. De tre studiene er valgt fordi de godt illustrerer de poengene jeg ønsker å løfte frem, ikke fordi de utmerker seg som spesielt graverende tilfeller.

Eksempel 1: Gruppevis feilrate
Leedham undersøker frekvensen av 14 kohesive uttrykk (vist i tabell 2) i engelske tekster som er skrevet av studenter med kinesisk som L1 (2011:199–228). Leedham sammenligner frekvensen av de 14 uttrykkene hos studenter som går i første og andre år på universitetet, med studenter som går i tredje år. Tallene er frekvensene av hvert uttrykk per millioner ord, så de er sammenlignbare uavhengig av tekstlengder. Leedhams analyse består i å sammenligne frekvensene parvis for hvert av uttrykkene, som vist i tabellen. For eksempel sammenlignes frekvensen av 'on the other hand' blant de yngre studentene (257) med frekvensen av 'on the other hand' blant de eldre studentene (129), mens

Tabell 2: Frekvenser av et utvalg kohesjonsuttrykk i tekster skrevet av kinesiske studenter i henholdsvis 1. eller 2. og 3. studieår. Frekvenstallene er gjengitt som antall forekomster per millioner ord. Forenklet gjengivelse etter Leedham (2011:303).

Kohesjonsuttrykk	**Signifikans**	**1. og 2. studieår**	**3. studieår**
on the other hand	*	257	129
besides	*	228	122
at the same time	**	185	65
nevertheless		157	179
nowadays		150	86
in the long run	**	114	22
and so on		114	50
in other words		100	108
at that time	**	86	14
meanwhile		86	72
what's more		64	29
last but not least		50	22
however		1290	1536
therefore	*	1140	868

frekvensen av 'besides' blant de yngre (228) sammenlignes med frekvensen av 'besides' blant de eldre (122), osv.

Til sammen innebærer dette at det er foretatt 14 slike parvise sammenligninger. Det er satt stjerne ved uttrykkene der hun har funnet det hun regner som signifikante forskjeller mellom de to gruppene, nærmere bestemt at $p < 0.05$. To stjerner markerer tilfeller der resultatene er "sterkt" signifikante, altså at *p*-verdiene er spesielt lave, i denne studien vil det si $p < 0.01$.[3] Konklusjonen er dermed at det finnes 6 statistisk signifikante forskjeller mellom gruppene, og at disse består i at hvert av de 6 aktuelle uttrykkene har høyere frekvens hos studenter i de lavere årskullene.

Dessverre er det et alvorlig problem ved denne fremgangsmåten. Når vi utfører en hypotesetest med α-nivå satt til 0.05, aksepterer vi en risiko

[3] Det går ikke eksplisitt frem av Leedhams fremstilling hvilken statistisk hypotesetest hun har brukt, men ut fra hva hun ellers skriver i avhandlingen, er det ikke usannsynlig at det er enten *t*-tester eller *z*-tester; for poenget mitt spiller ikke dette noen rolle.

på 5 % for en type-1-feil, altså for at vi feilaktig forkaster en sann nullhypotese. Dette kaller vi gjerne et falskt positivt resultat. Det er en konvensjon, og en form for kontrakt med leseren av forskningsartikkelen vår, at det finnes en risiko av denne størrelsen for et falskt positivt resultat. Men når vi utfører flere tester samtidig, er altså risikoen for et falskt positivt resultat 5 % for hver test. Det medfører at den samlede risikoen for minst ett falskt positivt resultat, det vi kaller gruppevis feilrate (*familywise error rate, FWER*), øker med antall tester. Hvis vi utfører 14 samtidige tester, som her, er den totale risikoen for minst ett falskt positivt resultat blitt over 50 % (Jensen 2018:460–461).[4] En så høy risiko for falske resultater er ikke del av konvensjonen og neppe i tråd med kontrakten med leseren.

Det mest grunnleggende problemet i fremgangsmåten til Leedham er at det blir testet på flere hypoteser uten at α-nivået blir justert for å motvirke gruppevis feilrate, altså den økte akkumulerte faren for type-1-feil eller falske positive funn. Det er dette Tukey kaller "the problem of multiplicity", som han i 1980 hevder blir sett bort fra i stor grad (Tukey 1980:24). Jeg tror ikke dette er blitt så mye bedre siden den gang, i hvert fall ikke i språkforskningen. Gries (2015:16–17) peker på det samme.

En vanlig løsning på problemet med økt gruppevis feilrate er å korrigere for den økte risikoen for feil. Bonferroni-korrigering blir mye brukt til dette formålet, eventuelt Šidák-korrigering, som er å foretrekke, særlig dersom antall sammenligninger er høyt (Abdi 2007).[5] Med slik korrigering ville α-nivået måtte settes til omtrent 0.005, og flere av de rapporterte positive funnene til Leedham ville vise seg ikke å være signifikante likevel.

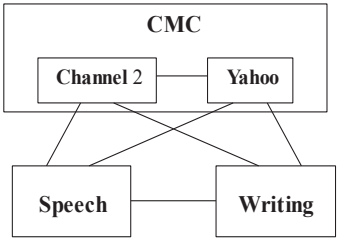
Problemet med å bruke så mange tester i en studie og korrigere for multippel testing eller gruppevis feilrate på denne måten er at risikoen for type-2-feil øker, altså risikoen for falske *negative* resultater. Ved å operere med et α-nivå på ca. 0.005 for hver enkelt test, slik man ville måtte gjøre i dette tilfellet, gir man altså avkall på styrke til å avdekke

[4] Dette kan regnes ut med formelen $1 – (1–\alpha)^k$, der k er antall samtidige tester. $1 – (1–\alpha)^{14} \approx 0.51$ for $\alpha = 0.05$.

[5] Bonferroni-korrigering er egentlig en unøyaktig tilnærming til Šidák-korrigering, utviklet fordi Šidák-korrigering er krevende å regne ut for hånd. Med dagens datamaskiner er ikke dét en viktig begrunnelse, og jeg anbefaler å bruke Šidák-korrigering hvis antall tester overstiger 3, for å unngå et unødvendig konservativt α-nivå.

tendenser som faktisk finnes i populasjonen. En annen måte å si det på er at med samme utvalgsstørrelse vil kravet til effektstørrelsen være strengere for å oppnå signifikant resultat *i hvert enkelt tilfelle*. Dette er en viktig grunn til at man bør begrense antall tester til et lite sett av velbegrunnede hypoteser. Jeg vet ikke om Leedham hadde en konkret, begrunnet hypotese for hvert enkelt av de 14 uttrykkene som ble testet, men det er ingenting i avhandlingsteksten som tyder på det, og det er lite trolig at hun hadde det. Hun skriver om konnektivene generelt og noen undergrupper av dem mer spesielt, men har ingen eksplisitt begrunnelse for at hver enkelt av dem skulle ha ulik frekvens i de to gruppene. Jeg mistenker med andre ord at hun bruker hypotesetester, men ikke til testing av hypoteser. Så en mer grunnleggende løsning på dette problemet ville være å arbeide mer for å utvikle et mindre antall konkrete og begrunnede hypoteser. En mulig fremgangsmåte ville være å benytte analysene fra 2011 som en utforskende analyse som kunne danne grunnlag for å generere spesifikke hypoteser om bare noen av konnektivene. Disse færre hypotesene kunne så testes på nye data, og analysen fra 2011 ville dermed ha fungert som en pilotstudie.

Eksempel 2: Anova-modeller og gruppevis feilrate
Det neste eksemplet er en studie i en avhandling av Nishimura (2008), og den har et lignende, men ikke identisk problem.



Figur 4: Diagram som viser sammenligningen av frekvenstall mellom fire ulike teksttyper. Gjengitt etter Nishimura (2008:120).

Nishimura analyserer frekvens av visse grammatiske trekk i fire teksttyper, nemlig muntlige, skriftlige og to typer av digitale skriftlige tekster eller datamaskinmediert kommunikasjon (*computer-mediated communication, CMC*). Hun sammenligner frekvensverdiene ved å utføre parvise

tester mellom hver av de fire teksttypene, totalt 6 tester, illustrert med strekene mellom boksene i diagrammet i figur 4. Dette skaper et tilsvarende problem som det vi nettopp så på, knyttet til multippel testing og økt gruppevis feilrate, men i dette tilfellet finnes det en standardløsning som er i vanlig bruk og anbefalt i lærebøker beregnet på lingvister (Gries 2013:276–280; Larson-Hall 2010:290–311; Levshina 2015:182–189; Lowie og Seton 2013:63). I stedet for å bruke *t*-tester for hver parvise sammenligning[6] er det i slike situasjoner vanlig heller å modellere hele systemet med variasjonsanalyse, ofte kjent som anova (*analysis of variance*). Anova tar hånd om sammenligningen av alle de fire gruppene på en gang innenfor én modell. Deretter er det vanlig å bruke en prosedyre for trinnvis reduksjon av modellen til man får det vi kaller en minimal adekvat modell, altså den minste – enkleste – modellen som beskriver dataene på en adekvat måte. Den minimale adekvate modellen blir tilegnet en *p*-verdi av statistikkprogrammet, som det er vanlig å rapportere og bruke til å avgjøre om modellen er "signifikant", slik jeg selv gjør det (Jensen 2017: for eksempel 96–97). I noen tilfeller vil den minimale adekvate modellen være identisk med den maksimale modellen, altså ved at ingen trinnvis reduksjon kunne utføres uten for stort tap av informasjon.

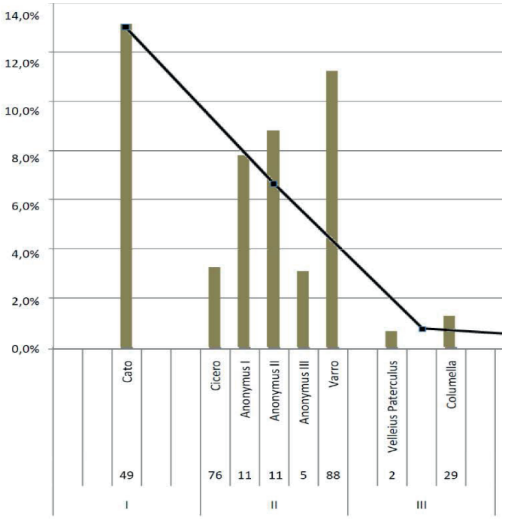
Problemet med denne mye brukte og allment aksepterte fremgangsmåten er at man ikke lenger egentlig tester noen spesifikk hypotese. Hypotesen til anova er at det finnes *en eller annen forskjell* i systemet av sammenligninger, men dette er lite spesifikt og i liten grad noe jeg ville anse for en egentlig hypotese. Særlig dersom flere enn én forklaringsvariabel er involvert og analysen inkluderer interaksjoner mellom forklaringsvariablene, kan systemet bli temmelig komplekst og "hypotesen" som testes, tilsvarende ullen. Etter hvert som man justerer anova-modellen gjennom trinnvis reduksjon, foretar man dessuten i realiteten en seleksjon mellom alternative modeller, og dermed justerer man også stadig hypotesen (og nullhypotesen) som blir testet. Som hypotesetesting betraktet er denne prosedyren ikke gyldig, og resultatene bør ikke betegnes som "signifikante".

I stedet er det naturlig å se på anova-modellering som en utforskende tilnærming til dataene. En slik tilnærming kan gi oss et overblikk og en

[6] I stedet for *t*-tester bruker Nishimura parvise kjikvadrat-tester på en måte som strider mot prinsippet om uavhengige observasjoner, så analysene hennes er ugyldige også av den grunn. Se neste seksjon.

forståelse av materialet som vi i neste omgang om ønskelig kan benytte til å utforme en mer spesifikk og testbar hypotese som da ville ha empirisk forankring i enkelte av tendensene i materialet og være knyttet til bare noen av strekene i diagrammet. Én mulig slik hypotese i Nishimuras studie kunne for eksempel være at digital skriving skiller seg både fra muntlig språk og fra tradisjonelt skriftlig språk. Deretter kunne man samle inn et nytt, tilsvarende materiale, og den nyutviklede hypotesen kunne så testes på det nye materialet ved hjelp av to *t*-tester (og Šidák-korrigering).

Eksempel 3: Uavhengige observasjoner
Det siste eksemplet er fra et litt annet felt; det dreier seg om en studie av diakron utvikling av bruken av en viss type underordning i latin, hentet fra en doktoravhandling av Danckaert (2011).



Figur 5: Søylediagram som gjengir frekvensen av en viss type underordningskonstruksjoner i 8 tekster skrevet på latin gjennom en periode på ca. 300 år (x-aksen). Figuren er gjengitt etter Danckaert (2011:303), men beskåret slik at den bare viser den delen som er relevant for min diskusjon her.

Danckaert (2011:300–304) bruker 8 latinske tekster fra 3 ulike perioder som strekker seg over til sammen 300 år, til å sammenligne frekvensen av den underordningskonstruksjonen han er interessert i, mellom disse tre periodene. Frekvensen blir i studien regnet ut som antall forekomster per antall mulige kontekster for den aktuelle konstruksjonen. Han finner at frekvensen i tekstene går ned over disse 3 periodene, og han bruker denne tendensen til å konkludere med at frekvensen går ned i latin generelt i denne perioden.

Problemet med denne analysen er at den er basert på ikke-uavhengige observasjoner (Jensen 2018:461–463); i den første perioden teller Danckaert 49 forekomster i i alt 372 potensielle kontekster i én tekst skrevet av én forfatter (Cato) og bruker disse observasjonene som en indikasjon på bruken i latin generelt i denne perioden. Han putter tallene inn i en test for sammenligning med de andre periodene, som om hver observasjon kom fra et tilfeldig utvalg av 372 tekster skrevet av 372 ulike personer.[7] Problemet gjelder også i de andre to periodene; det er flere observasjoner av samme tekst, og disse observasjonene er dermed ikke uavhengige av hverandre. Fremgangsmåten tilsvarer at jeg skulle ha stoppet samme mann 5 ganger på Torgallmenningen for å høre om han skarret hver gang, i stedet for å vurdere 5 forskjellige informanter. En slik fremgangsmåte lurer den statistiske testen til å tro at den har større sikkerhet for beregningene sine enn den i realiteten har, og slik blir det lettere å forkaste nullhypotesen og dermed oppnå et "signifikant" resultat. Det reelle totale utvalget til Danckaert er på bare 8 personer, fordelt på 3 perioder; én av periodene er representert ved bare én person. Det betyr at datamaterialet rett og slett er for lite til å kunne testes statistisk med tanke på å konkludere om latinsk språkbruk i de tre periodene generelt.

Hypotesen til Danckaert er derfor rett og slett ikke testbar, i hvert fall ikke med det datamaterialet som han har til rådighet. Et vanlig råd til forskeren i en slik situasjon ville være å gå tilbake og samle inn data fra flere informanter, men i og med at disse tekstene er rundt 2000 år gamle, så kan nok det være forbundet med vanskeligheter. Da er kanskje den eneste løsningen heller å gjøre en annen type studie, kanskje en kvalitativ analyse eller deskriptiv presentasjon. Andre alternativer kunne være å

[7] Danckaert bruker Wald-intervaller på binomialfordelinger for å teste disse forskjellene. Se f.eks. Scherer (2018). Hvorvidt han gjør dette eller bruker kjikvadrattest eller Fishers eksakte test, er uviktig for diskusjonen her; prinsippet om uavhengige observasjoner er det samme uansett hvilken test som blir brukt.

endre forskningsspørsmålet eller forkaste studien, og eventuelt rapportere fra den som en metodologisk mislykket tilnærming. Om man har tilgang til et større utvalg, kunne man utnytte en slik undersøkelse til hypotesedanning, som i de foregående eksemplene, men i dette tilfellet er det altså kanskje ikke mulig. Eksemplet viser at det i noen tilfeller rett og slett kan være umulig å teste en hypotese statistisk, og da er faktisk det eneste alternativet å la det være.

Utforskende tilnærming, ikke teknikker

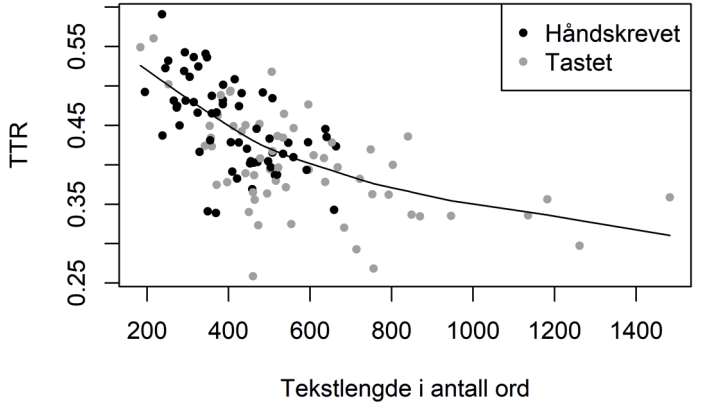
I dette kapitlet skal vi gå videre til de utforskende tilnærmingene og se nærmere på dem. Tukey understreker at man bør se på utforskende dataanalyse som en måte å *tilnærme* seg analysen på heller enn et sett med *teknikker*, slik hypotesetesting kan sies å være. "Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques" (Tukey 1980:23, Tukeys utheving). Det finnes ikke faste oppskrifter for utforsking. Man må tilpasse seg landskapet eller terrenget etter hvert som man oppdager det. Dette innebærer ikke at man ikke benytter teknikker i utforskende dataanalyse, men at det er kreativiteten og fleksibiliteten i utnyttelsen av disse teknikkene som er det essensielle. Vi skal se på noen teknikker som er til rådighet.

Visualisering

Visualisering og inspeksjon av visualiseringer står sentralt i det å utforske ukjente data. Man kan bruke enkle diagrammer som spredningsdiagram, boksdiagram, histogram og tetthetskurver, skreddiagram og vekstdiagram til å visualisere og inspisere variablenes egenskaper.

Spredningsdiagrammer (figur 6) viser typisk sammenhengen mellom to kontinuerlige variabler. I et spredningsdiagram representerer hvert punkt ett observert individ, men verdiene til to kontinuerlige variabler på dette samme individet samtidig. Såfremt antall observasjonsenheter ikke er for stort, kan man dessuten enkelt supplere med fargekoding av to eller flere verdier for en kategorisk variabel. Spredningsdiagrammer er godt egnet til å få frem korrelasjon (sammenheng) mellom to variabler på de samme individene, og diagrammet i figur 6 illustrerer tydelig en negativ korrelasjon mellom TTR for ordformer og tekstlengde; høye verdier av tekstlengde henger sammen med lave verdier av TTR og

motsatt. Dessuten går det tydelig frem at sammenhengen ikke er lineær; etter hvert som tekstene blir lengre, synker TTR-verdiene mindre bratt. Spredningsdiagrammer har i tillegg den fordelen at hver variabelverdi blir vist nøyaktig, slik at man lett kan peke ut blant annet maksimums- og minimumsverdier, og når antall observasjonsenheter ikke er for stort, kan man studere enkeltindivider i diagrammet. I diagrammet i figur 6 kommer det for eksempel frem at alle de lengste tekstene er tastet, mens det ikke er bare håndskrevne tekster blant de korteste tekstene.

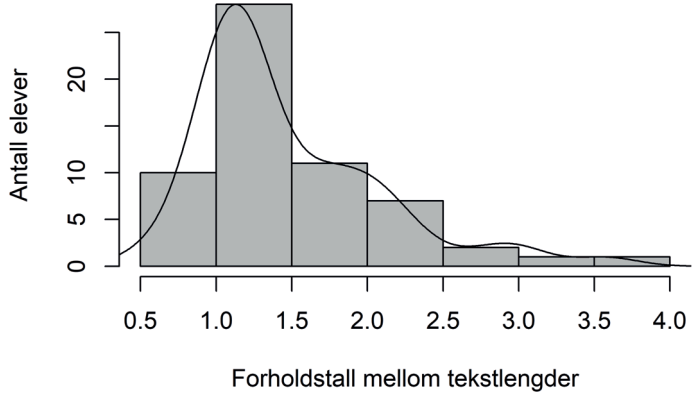


Figur 6: Spredningsdiagram for to kontinuerlige variabler som representerer teksters egenskaper. Eksempel gjengitt etter Jensen (2017:173). X-aksen representerer tekstlengde i antall ord, mens y-aksen representerer TTR for ordformer i teksten. Hvert punkt representerer én tekst, og en dikotom variabel (skriveverktøy) er gjengitt med fargen på punktene. Det er dessuten tegnet inn en ikke-parametrisk regresjonskurve som tydelig illustrerer den negative og ikke-lineære korrelasjonen mellom de to variablene.

Boksdiagrammer (figur 2) viser typisk sammenheng mellom en kontinuerlig variabel og en eller flere kategoriske variabler og er særlig egnet til å visualisere både den kontinuerlige variabelens fordeling og forholdet mellom ulike grupper. Boksdiagrammer er i sin vanligste form ikke-parametriske; medianen er representert ved den fete linjen midt i

boksen, mens boksens utstrekning markerer den øvre og nedre kvartil.[8] Slik visualiserer boksens utstrekning omfanget av den midterste, typiske halvparten av observasjonene. Minimums- og maksimumsverdier er markert ved tverrstrekene ytterst på halene på hver side av boksen, og i noen tilfeller er utliggere tegnet inn spesielt. I figur 2 er det to kategoriske variabler.

Histogrammer og nært beslektede tetthetskurver (figur 7) viser fordeling i én kontinuerlig variabel over et utvalg av individer – elever i dette tilfellet. Histogrammet deler inn verdiområdet i segmenter, typisk av lik bredde, og viser antall observasjoner i hvert segment som høyden på søylene. Tetthetskurven bruker utvalget som grunnlag for et estimat på en idealisert fordeling av uendelig mange observasjoner i populasjonen. Disse diagramtypene er særlig egnet til å få frem skjevheter og modusverdi, men de er dårlig egnet til å visualisere nøyaktige minimums- og maksimumsverdier i et utvalg.

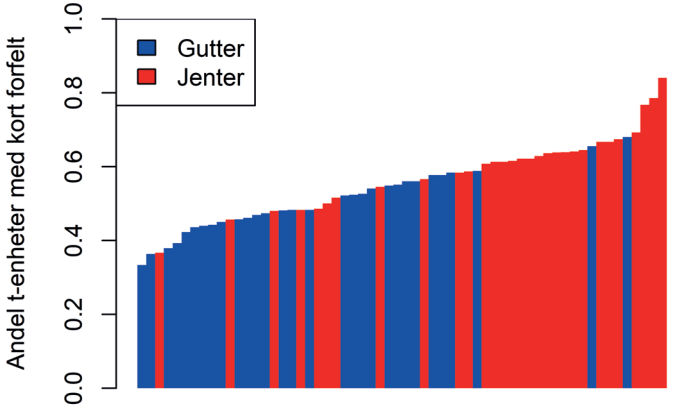


Figur 7: Histogram og tetthetskurve for en kontinuerlig variabel. Eksempeldiagram gjengitt etter Jensen (2017:97). Diagrammet viser at verdier mellom 1.0 og 1.5 er mest vanlig, og at fordelingen er tydelig høyreskjev.

Skreddiagrammer (figur 8) brukes i likhet med histogrammer til å visualisere fordelingen av en enkelt kontinuerlig variabel. I skred-

[8] Medianen er midtpunktet i en fordeling av variabelverdier, slik at det er like mange verdier over som under medianverdien. Tilsvarende deler kvartilene materialet inn i fire like store grupper.

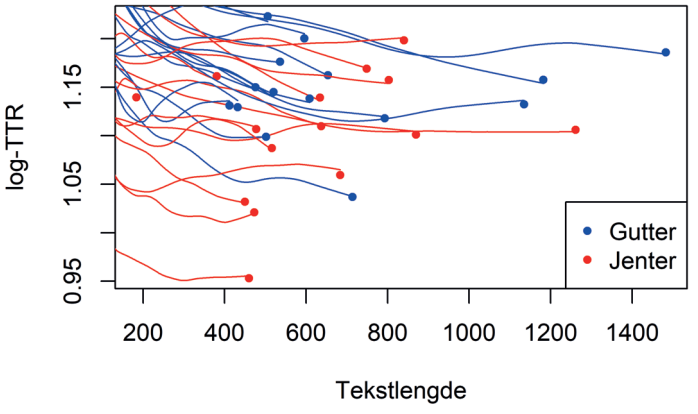
diagrammer er enkeltobservasjonene sortert etter verdi og typisk tegnet enten som punkter eller søyler, og diagramtypen er best egnet når antall observasjoner ikke er for stort. Ved hjelp av fargekoding er skreddiagrammer dessuten godt egnet til å få frem rangeringsegenskaper i interaksjon med en kategorisk variabel, som for eksempel kjønn. Skreddiagrammer er særlig egnet til utforsking av et materiale med tanke på utliggere eller atypiske enkeltindivider, for eksempel de to guttene med høye verdier i figuren, som man i etterkant kunne velge å gjøre kvalitative analyser av.



Figur 8: Skreddiagram over en kontinuerlig variabel, med fargekoding av en dikotom variabel (kjønn). Eksempel gjengitt etter Jensen (2017:292). Y-aksen viser andel t-enheter med kun ett ord i forfeltet i tekster skrevet med tekstbehandlingsverktøy av gutter og jenter. Diagrammet illustrerer at mange jenter bruker en stor andel korte forfelt i sine tekster.

Enkle diagramtyper kan kombineres til sammensatte diagrammer, f.eks. som i vekstdiagrammet i figur 9. Diagrammet bruker punkter til å vise verdier av en variabel, i dette tilfellet en logaritmetransformert variant av TTR for lemmaformer. Men i tillegg bruker diagrammet kurver til å vise hvordan disse TTR-verdiene utvikler seg gjennom hver enkelt teksts forløp. På denne måten kan man lettere få øye på om en variabel oppfører seg uvanlig i visse partier av teksten, og slik gjennom kvalitativ analyse utvikle bedre forståelse av hva slags tekstlige egenskaper variabelen representerer. I eksemplet i figur 9 blir det klart at verdien for log-

TTR *synker* utover i teksten i de fleste tilfellene, og diagrammet avslørte dermed at variabelen *ikke* er et tekstlengdeuavhengig mål for leksikalsk variasjon.



Figur 9: Vekstdiagram som viser verdier for en logaritmetransformert variant av TTR for lemmaformer i tekster skrevet med tekstbehandlingsverktøy (ett punkt for hver tekst), i tillegg til hvordan disse TTR-verdiene utvikler seg utover i teksten (én kurve for hver tekst). Kurvene er glattet med en matematisk funksjon, for å gi bedre overblikk. Fargene representerer skribentenes kjønn. Eksempel gjengitt etter Jensen (2017:194).

Alle diagramtypene ovenfor er enkle, men nyttige verktøy som rutinemessig kan brukes i det man gjerne kaller initiell eller innledende dataanalyse, når forskeren gjør seg kjent med variablers egenskaper for å avgjøre hva slags metoder som kan egne seg for videre analyse, og om fordelingene tilfredsstiller de matematiske forutsetningene for disse metodene.[9] Visualisering står sentralt i utforskende tilnærminger, men diagrammer er også godt egnet for deskriptiv presentasjon av data i ferdige forskningspublikasjoner.

Multivariate metoder

Det mange trolig først og fremst tenker på som utforskende tilnærminger, er mer avanserte multivariate analysemetoder. Prinsipalkom-

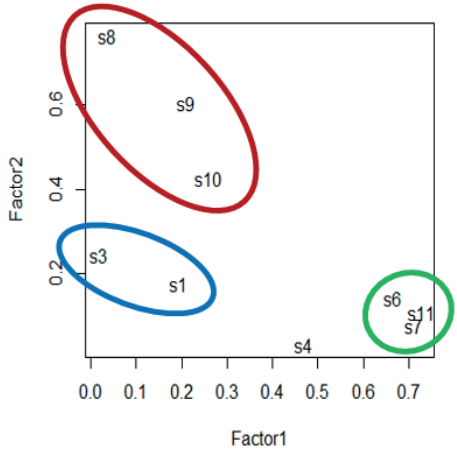
[9] Matematiske forutsetninger som ulike statistiske analysemetoder hviler på, er et stort og viktig felt som denne artikkelen i liten grad tematiserer.

ponentanalyse er alt presentert som et eksempel ovenfor, og vi skal nå i tillegg se på faktoranalyse og diskriminantanalyse. Disse tre analysemetodene er alle mye brukt i språkforskning, og de representerer tre ulike tilnærminger til komplekse datasett.

Prinsipalkomponentanalyse reduserer altså kompleksiteten i et mangedimensjonalt rom ved å samle så mye variasjon som mulig i få dimensjoner rent geometrisk. En teknikk som minner om dette, er utforskende faktoranalyse. Faktoranalysen tar også utgangspunkt i et mangedimensjonalt rom av flere variabler og prøver å redusere kompleksiteten i dette rommet, men i motsetning til prinsipalkomponentanalysen bygger faktoranalysen på et premiss om at det finnes latente, underliggende egenskaper eller *faktorer* som påvirker de enkelte observerbare variablene, og som dermed fremkommer i materialet gjennom korrelasjoner mellom disse variablene. Faktoranalysen har som mål å avdekke disse faktorene og beskrive dem som korrelasjoner med de observerbare variablene. Resultatet av en vellykket faktoranalyse kan derfor gjerne være lettere å tolke som en modell av et sett av teoretisk relevante fenomener, mens resultatet av en prinsipalkomponentanalyse består av geometriske dimensjoner som gjerne har mindre klar forbindelse til de opprinnelige variablene og kan være vanskeligere å gi en konseptuell fortolkning. På den annen side blir faktoranalyse kritisert for å være mer sårbar for påvirkning fra forskeren enn andre multivariate metoder (Jenset og McGillivray 2012:6). For eksempel må forskeren velge antall underliggende faktorer som analysen skal frembringe. Denne type sårbarhet kompliserer åpenbart både fortolkning av resultatene og spørsmål om validitet. Faktoranalyse har også noen premisser knyttet til variablenes egenskaper som prinsipalkomponentanalyse ikke har.

Eksemplet i figur 10 er fra en undersøkelse der respondenter vurderer stemmer og bilder av ansikter i en masketest etter forskjellige ferdigspesifiserte egenskaper vist i figur 10, som *hyggelig*, *ærlig*, *kjedelig*[10], osv. (Røyneland og Jensen 2020). En utforskende faktoranalyse finner at åtte av disse ni egenskapene grupperer seg i tre underliggende faktorer som vist i diagrammet (Røyneland 2018). Det vil si, egentlig fremkommer bare de to første faktorene, i grønt og rødt, i det todimensjonale diagrammet, men den blå grupperingen fremstår som en selvstendig

[10] Fordi *kjedelig* er det eneste adjektivet i listen med negativ verdi, er variabelen vendt i analysen; derfor står det *ikke kjedelig* i variabeloversikten.



s1	Hyggelig
s3	Ærlig
s4	Ikke kjedelig
s6	Selvsikker
s7	Sosial
s8	Intelligent
s9	Målrettet
s10	Selvstendig
s11	Kul

Figur 10: Et spredningsdiagram som viser ni variabler i de to første dimensjonene i en faktoranalyse. Ringene er tegnet inn manuelt på bakgrunn av resultatene av faktoranalysen. Diagrammet er hentet fra analysene som ligger til grunn for inndeling i holdningsdimensjoner hos Røyneland og Jensen (2020).

faktor om man går videre inn i analysen.[11] Denne utforskende teknikken antyder altså at egenskaper som *selvsikker*, *sosial* og *kul* kan oppfattes som uttrykk for den samme underliggende egenskapen av de ungdommene som har bidratt i undersøkelsen. Ved å samle disse variablene i én faktor, én egenskap, og gi denne en merkelapp (*dynamisme*) som indikerer egenskapens kjennetegn, gir analysen bedre oversikt over hvordan ungdommene vurderer de ulike maskene i undersøkelsen. Røyneland og Jensen bruker faktorene som fremkommer, som verktøy for å vurdere holdninger overfor personer med henholdsvis norsk og pakistansk utseende som snakker ulike dialekter. Røyneland og Jensen finner blant annet at unge respondenter fra Oslo vurderer ungdommer med Oslo-dialekt som mer *dynamiske* enn ungdommer med andre dialekter, uavhengig av norsk eller pakistansk utseende. Effekten gjelder imidlertid bare for gutter; jenter blir oppfattet som like dynamiske uavhengig av både utseende og dialekt.

[11] Selv om den blå faktoren fremkom gjennom faktoranalysen, ble den utelatt fra den endelige studien av Røyneland og Jensen på grunn av lavt antall variabler (2) og lav indre konsistens (Cronbachs $\alpha \approx 0.64$).

Det viser seg imidlertid å være en liten komplikasjon her. Dette er jo ikke holdningsvariabler som Røyneland og Jensen har funnet på. Det er variabler som er utviklet gjennom en lang linje av forskning (se referanser i Røyneland og Jensen 2020), og som man kjenner egenskapene til ganske godt. Tidligere forskning har utforsket nettopp hvordan disse ni variablene danner tre underliggende faktorer, men i analysen til Røyneland og Jensen viser det seg at to av variablene samvarierer med de andre variablene på en annen måte enn de har gjort i tidligere studier. I de tidligere studiene har *sosial* dannet en dimensjon sammen med *hyggelig* og *ærlig*, og ikke sammen med *selvsikker* og *kul*, mens *ikke kjedelig* har dannet en dimensjon sammen med *selvsikker* og *kul*.[12] I dette eksemplet ser vi dermed at også en teknikk som gjerne brukes utforskende, også kan brukes i et mer bekreftende – eller avkreftende – perspektiv, og det understreker at det ikke først og fremst er teknikken som er avgjørende, men tilnærmingen.[13]

Prinsipalkomponentanalysen har som nevnt ikke kjennskap til de tre teksttypene vi undersøkte, og på samme måte er faktoranalysen uvitende om teksttyper eller dialekter eller andre kategorier man forsøker å utforske. Disse teknikkene har i utgangspunktet tilgang bare til enkeltobservasjonene, altså de språklige variablene for hver tekst og de enkelte vurderingene av personlige egenskaper for hver maske i eksemplene ovenfor. De dimensjonene eller faktorene som analysene frembringer, kan deretter brukes som hjelpemiddel for å analysere for eksempel kategorier som teksttype eller dialekt, som forskeren er interessert i.

En *diskriminantanalyse* har derimot på en måte den motsatte tilnærmingen. I en studie kalt "Evaluation of texts in tests" (Golden, Kulbrandstad og Tenfjord 2017) undersøker forskerne tekster skrevet på norsk av personer med henholdsvis spansk og vietnamesisk som førstespråk. Tekstene er vurdert av sensorer i henhold til ferdighetsnivåene i *Det felles europeiske rammeverket for språk*, eller CEFR-nivåene (Europarådet og Utdanningsdirektoratet 2011). Disse nivåene går fra A1 (lavest) til C2 (høyest) og er formulert for norsk for eksempel i *Vurderingsskjema for norskprøven* fra Kompetanse Norge (2017). Ved

[12] De etablerte, tradisjonelle faktorene ser slik ut: *Status* = Intelligent, Målrettet, Selvstendig; *Attraktivitet* = Hyggelig, Sosial, Ærlig; *Dynamisme* = Kul, Ikke kjedelig, Selvsikker.

[13] Det finnes også varianter av faktoranalyse som kalles bekreftende (*confirmatory*) faktoranalyse og brukes spesifikt med dette formålet, men jeg går ikke inn på disse her.

å fortelle diskriminantanalysen hvilke av tekstene som er vurdert til A2, og hvilke som er vurdert til B1, og sammenholde denne informasjonen med tekstlige variabler som frekvensen av ulike feiltyper, gjennomsnittlig setningslengde og et mål for leksikalsk variasjon, kan diskriminantanalysen finne frem til hvilke av tekstenes egenskaper som er viktigst for sensorene som vurderer dem, eller i hvert fall hvilke egenskaper eller kombinasjoner av egenskaper som i størst grad kjennetegner de ulike nivåene. Med en prosedyre for trinnvis reduksjon av modellen, som ligner på den som ble beskrevet for anova-modeller ovenfor, står man til slutt igjen med en minimal, adekvat modell. I dette tilfellet fant Golden og kollegene hennes at det som spilte størst rolle for vurderingen, var tekstenes lengde pluss frekvensen av flere forskjellige feiltyper. Dette er selvfølgelig et resultat i seg selv, men trolig først og fremst en inngangsport til videre analyser.

Andre teknikker brukt utforskende

I tillegg til de tre multivariate metodene vi har sett kort på nå, har vi altså sett at også anova-modellering er mer aktuelt i en utforskende tilnærming enn som hypotesetesting. Det samme gjelder regresjonsanalyse, som anova kan ses som et spesialtilfelle av. Faktisk kan også teknikker som vi som regel omtaler som hypotesetester, brukes i et deskriptivt eller utforskende perspektiv (Amrheim, Trafimow og Greenland 2019). I omtalen av studien til Leedham ovenfor var jeg inne på at resultatene fra hennes 14 hypotesetester heller kunne fungere som en hypotesedannende pilotstudie, der de nye hypotesene i sin tur kunne testes på et nytt datamateriale. Med et slik formål er det bedre å tone ned *p*-verdiene fra testene og i stedet fokusere på effektmål og tilhørende konfidensintervaller (Cohen 1994; Gries 2005).

Effektmål er tall som indikerer styrken eller systematikken i en tendens, uavhengig av utvalgets størrelse. For forskjellen i middelverdi mellom to utvalg er Cohens *d* et praktisk og mye brukt effektmål; Cohens *d* angir forskjellen mellom middelverdiene målt i et vektet gjennomsnitt av utvalgenes standardavvik (Cohen 1988:20–27; Howell 2010:200; Torchiano 2018). For korrelasjonen mellom to variabler er Pearsons *r* et velkjent effektmål, der verdien 0 representerer ingen korrelasjon og 1 og –1 angir perfekt lineær sammenheng, henholdsvis positiv og negativ. For krysstabellanalyser er Cramérs *V* mye brukt. Når man angir effektmål basert på utvalg, er det alltid verdifullt samtidig å

angi et *konfidensintervall* for dette effektmålet. Et konfidensintervall beregnes på grunnlag av estimatet for utvalget kombinert med utvalgets størrelse og spredning og sier noe om hvor sikker man kan være på hva verdien av den reelle parameteren er i populasjonen. Det er vanlig å bruke 95 %-konfidensintervall, og dette angir da et spenn av verdier som med 95 % sannsynlighet dekker den sanne parameteren i populasjonen,[14] i dette tilfellet effektmålet. Konfidensintervallet er dermed et slags speilbilde av α-nivået på 5 %, og dersom begge ender av konfidensintervallet for Cohens *d* eller Pearsons *r* ligger på samme side av null, gjenspeiler dette et signifikant resultat.

Nytten av konfidensintervaller kan illustreres med studien fra Torgallmenningen, der estimatet fra utvalget var 0. En naiv tilnærming kunne hevde at studien viser at sannsynligheten er 0 for å påtreffe en ikke-skarrer i Bergen, eller at andelen av personer i Bergen som ikke skarrer, er 0. Dette vet vi jo ikke er korrekt, og det ville uansett være absurd å hevde noe slikt basert på et utvalg på bare 5 personer. En beregning av konfidensintervall for denne verdien gir som resultat at vi kan være 95 % sikre på at den reelle parameteren, altså andelen av ikke-skarrere i Bergen, ligger mellom 0.0 og 0.45. Dette er et ganske bredt spenn, og det illustrerer den store usikkerheten som følger med en undersøkelse basert på et så lite utvalg. Denne usikkerheten er det nyttig for leseren å få eksplisitt oppgitt.

I Leedhams studie kunne vi benytte *p*-verdiene fra de 14 testene til å danne nye, konkrete hypoteser kun for uttrykkene med $p < \alpha$, men uten å kalle disse signifikante i utgangspunktet. Eller vi kunne i stedet ha beregnet Cohens *d* med konfidensintervaller for hvert av de 14 uttrykkene med samme formål og formodentlig ganske likt resultat.

En annen grunn til å tone ned fokuset på *p*-verdier og testing av hypoteser er at det ofte egentlig ikke *er* noe vanskelig å oppnå signifikante resultater. Da Gosset utviklet *t*-testen, var det med tanke på å trekke konklusjoner fra analyser av små utvalg (Box 1987; Student 1908). Med store utvalg og stadig mer avanserte teknikker som i enda større grad er i stand til å sortere ut støy i utvalgene, er signifikante resultater ofte nærmest garantert og dermed i realiteten irrelevant. Dette er et poeng som har vært pekt på gjentatte ganger gjennom historien, men som fortsatt trenger å bli gjentatt (Bakan 1966; Berkson 1938; Gries

[14] Dette er ikke matematisk helt presist formulert, men det er praktisk å tenke på det på denne måten

2005; Kilgarriff 2005). Også i dette lyset er det mer meningsfullt å rapportere effektmål med konfidensintervaller enn signifikans og *p*-verdier.

Avslutning

Det har fremkommet en del kritikk av hypotesetesting (se f.eks. Wasserstein, Schirm og Lazar 2019), men kritikken rammer nok ikke først og fremst hypotesetesting som sådan, men hvordan disse analysene blir brukt. Jeg tror det er riktig å si at vi til tider opplever overdrevent fokus på signifikans og *p*-verdier. Kolleger forteller meg at de får artikkelmanus i retur fra tidsskriftsredaktører som etterlyser *p*-verdier. Dette kan være uheldig, når forskningsobjektet eller problemstillingen ikke er moden for hypotesetesting, eller kanskje rett og slett ikke egner seg for hypotesetesting. Det er også en innbakt fare i at tidsskrifter, konferanser og forskningsverden generelt er mer interessert i positive funn enn i negative (Nordahl 2019), altså at undersøkelser med lave *p*-verdier har større sjanse til å bli publisert enn undersøkelser med høye *p*-verdier. Det er viktig å ta inn over seg at også negative funn er funn, og dersom bare positive funn publiseres, skaper det et uriktig bilde av det aktuelle forskningsobjektet (van Hilten 2015), og vi vil se for mange ikke-replikerbare studier (men se også Amrheim et al. 2019). Med økt bruk av utforskende tilnærminger vil det kanskje bli enklere å få flere typer resultater inn i tidsskriftene.

Tittelen på denne artikkelen antyder at hypotesetesting og utforskende statistikk kan settes opp mot hverandre. Med artikkelen ønsker jeg å vise at dette er teknikker og perspektiver og tilnærminger som utfyller hverandre, og at en kombinasjon av tilnærminger ofte er en sunn tilnærming. Eller med Tukeys ord: "Analysis of data, with a more or less statistical flavour, should play many roles" (1980:24).

Referanser

Abdi, Hervé 2007. The Bonferonni and Šidák corrections for multiple comparisons. I N. Salkind (red.). *Encyclopedia of measurement and statistics*. Thousand Oaks: Sage.

Amrheim, Valentin, David Trafimow og Sander Greenland 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. The American Statistician, 2019, årg. 73 nr. S1, 262–270.

Bakan, David 1966. The test of significance in psychological research. Psychological Bulletin, 1966, årg. 66 nr. 6, 423–437.

Berkson, Joseph 1938. Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 1938, årg. 33 nr. 203, 526–536.

Biber, Douglas 1988. *Variation across speech and writing*. New York: Cambridge University Press.

Box, Joan Fisher 1987. Guinness, Gosset, Fisher, and small samples. Statistical Science, 1987, årg. 2 nr. 1, 45–52.

Chafe, Wallace L. og Jane Danielewicz 1987. Properties of spoken and written language. I R. Horowitz og S.J. Samuels (red.). *Comprehending oral and written language*. San Diego: Academic Press, 83–113.

Cohen, Jacob 1988. *Statistical power analysis for the behavioral sciences* (2. utg.). Hillsdale: Laurence Erlbaum.

Cohen, Jacob 1994. The earth is round ($p<.05$). American Psychologist, 1994, årg. 49 nr. 12, 997–1003.

Danckaert, Lieven 2011. *On the left periphery of Latin embedded clauses* (Doktoravhandling). Universiteit Gent, Gent. Hentet fra https://biblio.ugent.be/publication/1908358/file/4335555.pdf.

Europarådet og Utdanningsdirektoratet 2011. *Det felles europeiske rammeverket for språk: læring, undervisning, vurdering*. Oslo: Utdanningsdirektoratet.

Everitt, Brian S. og Torsten Hothorn 2011. *An introduction to applied multivariate analysis with R*. New York: Springer.

Golden, Anne, Lars Anders Kulbrandstad og Kari Tenfjord 2017. Evaluation of texts in tests, or: Where is the dog buried? I A. Golden, S. Jarvis og K. Tenfjord (red.). *Crosslinguistic influence and distinctive patterns of language learning: Findings and insights from a learner corpus*. Bristol: Multilingual Matters, 231–271.

Gries, Stefan T. 2005. Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff*. Corpus Linguistics and Linguistic Theory, 2005, årg. 1 nr. 2, 277–294.

Gries, Stefan T. 2013. *Statistics for linguists with R: a practical introduction* (2. utg.). Berlin: Mouton de Gruyter.

Gries, Stefan T. 2015. Quantitative designs and statistical techniques. I D. Biber og R. Reppen (red.). *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press, 50–72.

Gujord, Ann-Kristin Helland, Sandra Halverson og Bård Uri Jensen 2017. *Subjektrealisering i norsk og engelsk i ulike språklege modusar.* Innlegg presentert ved MONS 17, Os, Hordaland.

Halliday, Michael A.K. 1989. *Spoken and written language* (2. utg.). Oxford: Oxford University Press.

Horowitz, M.W. og A. Berkowitz 1964. Structural advantage of the mechanism of spoken expression as a factor in differences in spoken and written expression. Perceptual and motor skills, 1964, årg. 19, 619–625.

Howell, David C. 2010. *Statistical methods for psychology* (7. utg.). Belmont: Wadsworth.

Jensen, Bård Uri 2017. *Leksikosyntaktiske trekk og skriveverktøy: En kvantitativ undersøkelse av tekster skrevet for hånd og på tastatur av elever i VG1* (Doktoravhandling). Universitetet i Bergen, Bergen. Hentet fra http://bora.uib.no/handle/1956/16998.

Jensen, Bård Uri 2018. Er resultatet gyldig? Noen utfordringer ved bruk av kvantitative metoder i andrespråksforskning. I A.-K.H. Gujord og G.T. Randen (red.). *Norsk som andrespråk – perspektiver på læring og utvikling*. Oslo: Cappelen Damm Akademisk, 449–466.

Jenset, Gard B. og Barbara McGillivray 2012. Multivariate analyses of affix productivity in translated English. I M.P. Oakes og M. Ji (red.). *Quantitative methods in corpus-based translation studies: a practical guide to descriptive translation research*. Amsterdam: John Benjamins, 301–324.

Kilgarriff, Adam 2005. Language is never, ever, ever, random. Corpus Linguistics and Linguistic Theory, 2005, årg. 1 nr. 2, 263–276.

Kompetanse Norge 2017. *Vurderingsskjema for norskprøven: Delprøve i skriftlig framstilling*. Oslo: Kompetanse Norge.

Larson-Hall, Jenifer 2010. *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Leedham, Maria 2011. *A corpus-driven study of features of Chinese students' undergraduate writing in UK universities* (Doktoravhandling). The Open University. Hentet fra http://oro.open.ac.uk/29228/425/Maria_Leedham_OU_Phd_thesis_July_2011_post_viva.pdf.

Levshina, Natalia 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Lowie, Wander og Bregtje Seton 2013. *Essential statistics for applied linguistics*. Basingstoke: Palgrave Macmillan.

Nishimura, Yukiko 2008. *Aspects of Japanese Computer-Mediated Communication: Linguistic and Socio-Cultural Perspectives* (Doktoravhandling). Sheffield Hallam University. Hentet fra http://shura.shu.ac.uk/3212/2/10697431.pdf.

Nordahl, Marianne 2019. Forskere vred på resultatene i artikler om hjerteforskning. Hentet 07.03.2020 fra https://forskning.no/etikk-forskningsetikk-medisin-og-helse/ny-studie-forskere-vred-pa-resultatene-i-artikler-om-hjerteforskning/1337042.

Røyneland, Unn 2018. *What should you sound like to sound like you belong? A visual-verbal guise study of attitudes towards dialect and 'standard' in Norway*. Innlegg presentert ved Standard Languages in Europe: Attitudes & Perception, University of Vienna.

Røyneland, Unn og Bård Uri Jensen 2020. Dialect acquisition and migration in Norway: Questions of authenticity, belonging and legitimacy. Journal of Multilingual and multicultural development, 2020. https://doi.org/10.1080/01434632.2020.1722679.

Scherer, Ralph 2018. PropCIs: Various confidence interval methods for proportions. R package version 0.3-0. Hentet fra https://CRAN.R-project.org/package=PropCIs.

Student 1908. The probable error of a mean. Biometrika, 1908, årg. 6 nr. 1, 1–25.

Torchiano, Marco 2018. effsize: Efficient effect size computation: R package version 0.7.4. Hentet fra https://CRAN.R-project.org/package=effsize.

Tukey, John W. 1980. We need both exploratory and confirmatory. The American Statistician, 1980, årg. 34 nr. 1, 23–25.

van Hilten, Lucy Goodchild 2015. Why it's time to publish research "failures". Hentet 06.03.2020 fra https://www.elsevier.com/connect/scientists-we-want-your-negative-results-too.

Wasserstein, Ronald L., Allen L. Schirm og Nicole A. Lazar 2019. Moving to a world beyond "$p < 0.05$". The American Statistician, 2019, årg. 73 nr. sup1, 1–19. https://doi.org/10.1080/00031305.2019.1583913.

English summary

This article discusses the relationsship between confirmatory and exploratory statistical approaches in linguistic research. It uses examples from published studies to point out some challenges related to hypothesis testing, like the requirement for specific and explicit hypotheses, the assumption of independent observations and the principle of familywise error rate, and shows how some of these challenges may be met by adjusting the analytic approach. Further, it is being argued that the focus on statistical significance and *p*-values may have become too strong in some branches of linguistic research, and that reporting results in different ways and complementing with other analytic approaches might be advantageous. The article demonstrates a small bundle of visualisation techniques and emphasises their value as tools in an initial phase of analysis, and also explains at a superficial level three more advanced multivariate exploratory methods and their potential role as tools for hypothesis generation. The main point of the article is to show how confirmatory and exploratory approaches interact, and that both play crucial roles in linguistic research.

Keywords: *Quantitative analyses*; *hypothesis tests*; *explorative approaches*; *significance*; *visualisation; multivariate methods*