Novus forlag · eISSN 2387-6670

Flervalgstester i studier av metaforforståelse blant norskinnlærere[1]

Anne Golden[1], *Bård Uri Jensen*[2], *Oliwia Szymańska*[1].
[1] MultiLing, Universitetet i Oslo, [2] Høgskolen i Innlandet

Sammendrag

I denne artikkelen diskuterer vi ulike aspekter ved bruk av flervalgstest som metode i ordforrådsstudier med fokus på innlærere av norsk. Vi tar utgangspunkt i fire studier der denne metoden er brukt for å undersøke forståelse av norske metaforiske uttrykk, og drøfter ulike fenomen som framkommer i de ulike studiene. I tillegg presenterer vi noen betraktninger omkring statistisk analyse i slike studier og diskuterer mulige utfordringer som er knyttet til de valgene som tas. Fokuset er rettet hovedsakelig mot valg av uttrykk, utforming av distraktorer, kontekstens innvirkning og statistisk styrke. Siden fokuset vårt er på innlærere med et annet førstespråk enn norsk, diskuterer vi også hva ekvivalens mellom metaforiske uttrykk i ulike språk vil si, og bruker eksempler på forskjeller og likheter mellom norsk og polsk i denne sammenlikningen.

Nøkkelord: *flervalgstester, metaforer, metaforiske uttrykk, statistisk styrke, norsk som andrespråk, norsk som fremmedspråk, tverrspråklig ekvivalens*

[1] Dette arbeidet er delvis finansiert av Norges forskningsråd (NFR) gjennom ordningen for Sentre for fremragende forskning, prosjektnummer 223265.

Innledning

Ordforrådet i norsk – som i alle andre språk – er stort, og ordforrådsforskning er derfor krevende. Særlig gjelder dette når forståelsen skal studeres; det er viktig å holde produksjonsferdighetene utenfor slik at ikke deltakernes (mangel på) produktive ferdigheter påvirker resultatene. En hensiktsmessig metode kan derfor være en såkalt flervalgstest, der testdeltakerne krysser av for ett av flere alternativer som de mener dekker testordets betydning. En flervalgstest er et klassisk mål som er lett å administrere, rask å utføre og enkel å skåre. Dette bekrefter Prentice (2010) etter en pilotundersøkelse der hun prøvde ut ulike format til sin studie av metaforforståelse. Hun konkluderte med at flervalgsformatet egnet seg bedre enn å be elevene selv skrive hva et metaforisk uttrykk betyr, fordi elevenes egne svar ofte «på olika sätt kunde ligga nära den konventionaliserade, efterlysta betydelsen, utan att vara helt korrekta» (Prentice, 2010, s. 48). For små barn kan det være for kognitivt krevende å forklare og utdype ord og uttrykk. Mangel på gode forklaringer kan derfor gi misvisende resultat, siden barn kan forstå et uttrykk uten å kunne gi det en adekvat forklaring (Vosniadou, Ortony, Reynolds & Wilson, 1984). Imidlertid har flervalgstesten sine begrensninger. Den hevdes bare å vise testtakernes teststrategier (Stevenson, 2010; Jones, 2021) og si lite eller ingenting om graden av forståelse. Muligens måler den mer gjenkjennelse av et uttrykk (*recognition knowledge*) (Hughes, 2003, s. 76), men gjenkjennelse av et ord eller uttrykk *er* et steg mot det å forstå ordet eller uttrykket (Golden, 2014). Gode flervalgsoppgaver er riktignok «notoriously difficult to write» (Hughes, 2003, s. 3), men grundig planlegging og bevisst utforming av oppgavesettet hjelper betraktelig.

Gjennom eksempler fra fire norske studier går vi i denne artikkelen kritisk gjennom ulike sider ved bruk av flervalgstester. Disse studiene er Golden (2005), Golden og Larsen (2005), Golden og Szymańska (2021), og Szymańska og Golden (u. a.). Vi ønsker å vise at det er mange forhold å ta hensyn til, både i utvelgelsen og kategoriseringen av uttrykkene, utformingen av testen og analysen av svarene. De ulike metodiske valgene må imidlertid tilpasses formålet med studien for å etterstrebe konstruktvaliditet, dvs. at testen måler det den er ment å måle. I alle våre eksempler er fokuset på bruk av flervalgstest som metode for testing av metaforiske ord og uttrykk i norsk som andre- eller fremmed-

språk, og perspektivet er hovedsakelig knyttet til leseforståelse. Det er ulike syn på oppdeling av leseferdighet i komponenter, men mange studier viser at sammenhengen mellom ordforråd og leseforståelse er sterk (Sweet & Snow, 2003; Simmons, Rupley, Simmons & Graham, 2011; Ash & Baumann, 2017), og barn og unge som har lite utviklet ordforråd, har ofte vansker med å forstå det de leser (Kulbrandstad, 2018). Metaforer er intet unntak og viser seg å forekomme i alle teksttyper. I studiene vi introduserer, har forskerne en kognitiv tilnærming til metaforer i tråd med det erfaringsbaserte rammeverket (*experientialism*) som vi presenterer nedenfor. I dette rammeverket er én viktig variabel uttrykkets forekomst i studentenes førstespråk (L1) og for så vidt i hele deres språklige repertoar. Derfor ser vi også på tverrspråklige likheter og forskjeller. Siden det i ordforrådsstudier fort blir mange variabler som må veies opp mot hverandre, inkluderer vi i tillegg statistiske refleksjoner, særlig knyttet til statistisk styrke.

Artikkelen er bygget opp slik: Først presenterer vi flervalgstest som metode og gjennomgår fire norske studier som benytter flervalgstester for å undersøke flerspråklige elever og studenters metaforforståelse. Vi viser eksempler på oppgaver og peker på forskjeller i tilleggsspørsmål og i utvalget av deltagere. Siden det er forståelse av metaforiske ord og uttrykk som er i fokus, presenterer vi også kort vår tilnærming til metaforer her. Vi ser så på hva tverrspråklig sammenligning av ordforrådet innebærer, og bruker eksempler fra norsk og polsk i diskusjonen av likhet og ulikhet mellom uttrykk i forskjellige språk. Deretter introduserer vi grunnleggende teori om styrke i statistiske beregninger. I de påfølgende delene drøfter vi prinsipper og problemstillinger ved testdesign, datautvalg, analyse og ekvivalens mellom språk, med illustrasjoner fra de fire eksempelstudiene. Til slutt kommer en kort oppsummering.

Flervalgstester og metaforforståelse

Flervalgstester (*multiple choice*) er en av de vanligste måtene å teste ordforrådet på og har blitt brukt i standardiserte tester som TOEFL[2] gjennom en årrekke (Nation, 2001). Designet i en flervalgstest er – som be-

[2] *Test Of English as a Foreign Language*, en svært utbredt engelsktest som brukes over hele verden.

tegnelsen indikerer – en test hvor testdeltakeren har valg. De kommer i flere varianter, f.eks. med eller uten (innledende) kontekst for ordene som testes, med eller uten bruk av flere språk, med korte eller lange svaralternativer, osv. En flervalgstest som tester ordforrådet, består av en linje (*innledningen*) hvor testuttrykket forekommer med eller uten kontekst. Denne er etterfulgt av noen *svaralternativer*, som enten er *distraktorer* (dvs. alternativer som ikke stemmer) eller det *rette alternativet*, listet opp i tilfeldig rekkefølge, ofte under hverandre (se eksempel på neste side). Svaralternativene kan være synonymer, antonymer, forklaringer, eksempler på bruk, oversettelser m.m. For testing av ordforståelse skal testen vise om testdeltakerne vet hva testuttrykket betyr, og dette gjør deltakerne ved å velge ett av svaralternativene.

De fire studiene vi presenterer under, tester alle forståelsen av frekvente norske metaforiske ord og uttrykk. Selv om dette er en spesiell del av ordforrådet og krever litt ekstra når en tester flerspråklige personer (se delen om parallelle uttrykk i L1 og L2), er framgangsmåten den samme som ved testing av andre ord og uttrykk. Før vi presenterer flervalgsoppgavene, definerer vi imidlertid noen sentrale metaforbegreper fra det erfaringsbaserte rammeverket, som disse studiene bygger på. I dette rammeverket skiller en mellom *begrepsmetaforer* (også kalt *konseptuelle metaforer* eller *underliggende metaforer*) og *metaforiske uttrykk*. En begrepsmetafor defineres som en overføring mellom to domener i det konseptuelle systemet, og et metaforisk uttrykk er realiseringen av denne overføringen (Lakoff, 1993), eller for å si det på en enkel måte: begrepsmetaforen «er et tankekonsept som reflekteres i språket» (Askeland, 2019, s. 16). Med andre ord har vi et metaforisk uttrykk når et ord eller et uttrykk fra et konkret eller mer kroppsnært domene (kildedomenet) blir brukt for å uttrykke et abstrakt fenomen (i måldomenet). Når vi f.eks. vil formidle at noen har skiftet mening eller har ombestemt seg, kan vi uttrykke det som at noen har *snudd* i en sak. Den kroppslige handlingen *å snu* (seg) brukes altså for den abstrakte handlingen «å ombestemme seg». Når noen er blitt mer enige, kan dette uttrykkes som at de har *nærmet seg* hverandre. Både *snu* (i en sak) og *nærme seg* (hverandre) er metaforiske uttrykk, og de er realiseringer av den samme begrepsmetaforen ENIGHET ER SAMLOKALISERING.[3]

[3] I det erfaringsbaserte rammeverket er det vanlig å skrive begrepsmetaforer med store bokstaver på denne måten, dvs. X (fenomen i måldomene) ER Y (fenomen i kildedomene).

Oppgavesettet – som altså består av flervalgsoppgaver – i alle de fire eksempelstudiene ble laget i forbindelse med undersøkelsen som er presentert i Golden (2005), og det er brukt i sin helhet i Golden og Larsen (2005) og Golden og Szymańska (2021). I Szymańska og Golden (u. a.) er deler av dette settet brukt. Imidlertid er det lagt til ulike bakgrunns- eller tilleggsspørsmål i undersøkelsene som gjør dem ulike og interessante å diskutere. Det opprinnelige oppgavesettet ble laget med utgangspunkt i metaforiske uttrykk etter definisjonen ovenfor – både enkeltord og flerordsuttrykk[4] – som fantes i lærebøker i samfunnsfag på ungdomstrinnet. Disse uttrykkene ble valgt ut manuelt ved nærlesing av fire bøker og skumlesing av fem andre (se detaljer i Golden, 2005). De metaforiske uttrykkene som skulle testes, ble presentert i en kontekst, og elevene skulle velge mellom fire svaralternativer, dvs. ett rett svar og tre distraktorer. Først ble det gitt et eksempel som skulle forklares av læreren eller den som administrerte testen:

Å *slå alarm* om forurensning vil si

x	a.	å si fra at det er forurensning
	b.	å slå seg på grunn av forurensning
	c.	å ringe på grunn av forurensning
	d.	å bli forskrekket av forurensning

Testuttrykket var altså nevnt i innledningen, det ble kursivert, og sammen med riktig alternativ danner innledningen en setning som forklarer eller definerer testuttrykket. Innledningen består av et subjekt – som oftest en leddsetning eller en infinitivsfrase – og et verbal, og alternativene utgjør objektet eller predikativet i setningen. Den delen av innledningen som utgjør subjektet («å *slå alarm* om forurensning»), kalles også *oppgavestammen* i flervalgsoppgaven. Sammen med innledningen gir alternativene forskjellige typer forklaringer til testuttrykket eller uttrykker spørsmål og svar. Elevene fikk også tilsvarende oppgaver med ordene i oppgavestammen brukt med konkret betydning. En pilotversjon med 111 oppgaver ble først presentert for tre elever på 10. trinn med norsk som førstespråk, samme trinn som elevene i Golden (2005). Disse elevene ble intervjuet etter testen, og deres svar og synspunkter ble brukt både for å luke ut uheldige oppgaver og for å redusere

[4] Det er vanlig å kalle både ord og uttrykk brukt metaforisk for metaforiske *uttrykk*.

det endelige oppgavesettet til et antall oppgaver (81) som passet for en skoletime på 45 minutter.

I Golden (2005) ble det innhentet relevante bakgrunnsopplysninger som skulle brukes videre i analysen, bl.a. spørsmål om hvilke(t) hjemmespråk elevene brukte, om hvilke(t) språk de syntes de kunne best, om de hadde gått i norsk barnehage, og når de begynte på norsk skole. Til sammen deltok 400 ungdomsskoleelever. 170 av elevene hadde norsk som andrespråk (L2); i denne gruppen var over 20 ulike førstespråk representert. Resten av elevene utgjorde en kontrollgruppe med norsk som L1.

Golden og Larsen (2005) undersøkte forståelse av metaforiske uttrykk hos 79 grunnkurselever i videregående skole, hvorav 61 hadde norsk som L2. Metoden og oppgavene var de samme som i Golden (2005), men i denne undersøkelsen ble elevene i tillegg bedt om å gradere lytte-, snakke-, lese- og skriveferdighetene sine både på L1 og L2.

Golden og Szymańska har gjennomført to studier av polske studenters forståelse av metaforiske uttrykk. Hovedformålet med disse studiene er å undersøke metaforforståelse når man lærer norsk som fremmedspråk, og når konteksten norskopplæringen foregår i, er utenfor et norsktalende miljø. Disse studiene skiller seg fra Golden (2005) og Golden og Larsen (2005) ved at gruppene er langt mer homogene og noe eldre: alle innlærerne har samme førstespråk, de er rundt 20 år og har sannsynligvis omtrent samme motivasjon for å lære norsk. Deres språklige innputt kommer hovedsakelig fra lærebøkene de bruker, i tillegg til lærerens kunnskap og metaspråklige bevissthet, dvs. det læreren mener er viktig å vektlegge i undervisningen. Basert på pensumet og studieplanen skal studentene mestre norsk på tilnærmet B1- eller B2-nivå i henhold til *Det felles europeiske rammeverket for språk* (CEFR) (Utdanningsdirektoratet, 2011), avhengig av om de er andre- eller tredjeårsstudenter. I Golden og Szymańska (2021) deltok 5 andreårs- og 12 tredjeårsstudenter i norsk filologi ved et polsk universitet. Til forskjell fra Goldens (2005) studie ble de metaforiske uttrykkene delt i tre kategorier i analysen: a) uttrykk med fullstendig ekvivalens i polsk, dvs. den samme betydning formulert med tilsvarende uttrykk[5], b) uttrykk med

[5] Dette ble vurdert av erfarne språkbrukere; «tilsvarende uttrykk» vi si at de enkelte ordene tilsvarte hverandre i begge språkene ved at de viste til (omtrent) det samme meningsinnholdet.

delvis ekvivalens (den samme betydning, men med noe forskjellig formulering), c) uttrykk uten ekvivalens i polsk. Analysen hadde fokus på betydningen av ekvivalens mellom uttrykkene i L1 og L2, særlig på hvilken rolle delvis ekvivalens spilte.

I Szymańska og Golden (u. a.) er det færre uttrykk i oppgavesettet, og de er begrenset til realiseringer av de fire begrepsmetaforene Å FORSTÅ ER Å SE, ENIGHET ER SAMLOKALISERING, FØLELSER ER TEMPERATUR, ENHET ER POSITIVT. Deltakerne var 25 polske bachelorstudenter som alle studerte norsk filologi ved et polsk universitet, 14 var andreårsstudenter og 11 var tredjeårsstudenter. Som i de andre eksempelstudiene ble deltakerne også her testet for forståelse av den konkrete betydningen av ordene i de metaforiske uttrykkene. Men i denne studien ble det gjort gjennom selvrapportering, studentene fikk oppgitt ordene og skulle krysse av for ett av tre alternativer: «jeg forstår», «jeg tror jeg forstår» og «jeg forstår ikke». Totalt bestod testen av 28 oppgaver med metaforiske uttrykk og 25 oppgaver der uttrykkene var brukt konkret. I tillegg til å velge mellom alternative forklaringer til de metaforiske uttrykkene ble deltakerne bedt om å oppgi om de hadde gjettet. Analysen konsentrerte seg om betydningen av ekvivalensgrad, dvs. om det var fullstendig, delvis eller ingen overlapping mellom uttrykkene i innlærerens førstespråk og norsk. I tillegg ble forekomsten av uttrykk fra disse begrepsmetaforene i polsk også trukket inn.

I konstruksjonen av flervalgsoppgaver er både innholdet og strukturen i innledningen, antall distraktorer og den semantiske avstanden mellom de ulike svaralternativene i oppgavesettet viktige faktorer. Når forståelse av metaforiske uttrykk er temaet, og en bruker det erfaringsbaserte rammeverket (se over), kan også utvalg av uttrykk fra samme begrepsmetafor være relevant. I tillegg er det i studier med flerspråklige deltakere også interessant å diskutere betydningen av uttrykkets forekomst (og eventuelt hvor vanlig det er) i testtakernes førstespråk, samt graden av ekvivalens mellom uttrykkene i første- og andrespråket, når en analyserer svarene. I Golden (2005) ble også andre variabler analysert, som uttrykkets omfang, tema i innledningen, bildestyrke (hvor gjennomsiktig uttrykket var) og frekvensen (i norsk) av de ulike ordene i det metaforiske uttrykket.

Statistisk analyse[6]

Hvilke tilnærminger til statistisk analyse en velger å bruke, avhenger av datamaterialet man har tilgjengelig, og forskningsspørsmålene man ønsker svar på. Denne seksjonen presenterer noen teoretiske prinsipper for statistisk analyse, spesielt knyttet til statistisk styrke. I de fire påfølgende seksjoner blir statistiske tilnærminger i de fire eksempelstudiene drøftet, basert på prinsippene presentert her.

Utvalgsstørrelse, statistisk styrke og risiko for feilslutninger

Når man planlegger en undersøkelse med hypotesetesting, er størrelsen på utvalget et sentralt spørsmål. Spørsmålet har dessverre ingen klare svar, men henger sammen med vurderinger av risiko for å trekke feilaktige konklusjoner. Risikoen for *falske positive resultat* kontrollerer forskeren ved å velge et signifikansnivå α (*alfa*) å sammenlikne p-verdien[7] fra testen med. Risikoen for *falske negative resultat* er det imidlertid gjerne mindre oppmerksomhet rundt, og denne risikoen er sjelden nevnt eksplisitt i studier.

En hypotesetests statistiske styrke (*power*) – som kan benevnes π (*pi*) – indikerer sannsynligheten for å *forkaste* en falsk nullhypotese, altså sannsynligheten for å oppnå et statistisk signifikant resultat når hypotesen er sann (Larson-Hall, 2010, s. 104–114). Risikoen for å *bekrefte* en falsk nullhypotese kalles gjerne β (*beta*), og styrken blir dermed $\pi = 1 - \beta$. Styrken er avhengig av signifikansnivået α, utvalgsstørrelsen N og størrelsen på effekten i populasjonen[8]; når disse tre er kjent, kan styrken regnes ut (Cohen 1992a). Hvis man ønsker lavere risiko for falske positive resultat, kan man senke α, men det medfører økt risiko for falske negative resultat, altså β. Forholdet mellom falske positive og falske negative resultat er dermed en avveining eller et kompromiss, og en vanlig anbefaling fra statistikere er å sette $\alpha = 0{,}05$ (5 %) og $\beta = 0{,}20$

[6] For forklaring av grunnleggende statistiske begreper som er omtalt i denne seksjonen, se en lærebok i statistikk (f.eks. Field, Miles & Field, 2012; Larson-Hall, 2010; Levshina, 2015).

[7] *P-verdien* angir sannsynligheten for at man i et tilfeldig utvalg fra en populasjon der nullhypotesen er sann, vil finne en så sterk tendens som den man har funnet i utvalget.

[8] *Effekt* angir hvor systematisk en tendens er, uavhengig av størrelsen av et eventuelt utvalg. Vanlige effektmål er Cohens *d* og Pearsons *r*. Se f.eks. Larson-Hall (2010, s. 114–120) eller Jensen (2020, s. 65–66).

(20 %) (Pripp, 2017). Nødvendig utvalgsstørrelse for en studie avhenger dermed av hvor store effekter man ser etter, og dessuten hvilke risikoer for falske resultat man er villig til å akseptere. Forventninger om effektstørrelser i materialet kan være basert på tidligere studier av lignende data, eller man kan foreta en pilotstudie. Dette estimatet kan så danne grunnlag for beregning av hvor store utvalg som er nødvendig for å oppnå tilstrekkelig statistisk styrke i studien.

I andrespråksstudier er datainnsamling gjerne kostbar eller begrenset av tilgangen på tilstrekkelig antall passende informanter. Slike begrensninger kan gjøre det mindre aktuelt å gjennomføre pilotstudier av nødvendig størrelse. Om man dessuten mangler tidligere studier som er tilstrekkelig sammenlignbare, har man lite grunnlag for å anta noe om effektstørrelsen i det aktuelle materialet på forhånd. Uten et slikt estimat kan man i stedet ta utgangspunkt i hva slags effektstørrelse som kan anses for å ha faglig eller praktisk betydning for det aktuelle forskningsspørsmålet. Denne effektstørrelsen kan dermed legges til grunn for beregningen av nødvendig utvalgsstørrelse i studien. Som allerede nevnt, og som dessuten går frem av utvalgsstørrelsene i Golden (2005) og Golden og Larsen (2005), er flervalgsoppgaver en metode som er relativt lite kostnadskrevende å utføre i større skala, og som dermed i større grad gjør det mulig å gjennomføre pilotstudier, forutsatt tilstrekkelig tilgang på relevante deltakere. Med utgangspunkt i den antatte eller ønskede effektstørrelsen og valgt α og β kan man regne ut nødvendig utvalgsstørrelse. Vanlige statistikkprogram har funksjoner som gjør dette automatisk.

Tabell 1 viser hvilke effektstørrelser som kan avdekkes av ulike utvalgsstørrelser ved $\alpha = 0{,}05$ og $\beta = 0{,}20$. Den første raden gjelder testing av forskjeller i en variabel mellom to like store delutvalg, hver av størrelse N, der effekten er målt som Cohens *d*. Den andre raden gjelder korrelasjonstester mellom to variabler i et utvalg av størrelse N, der effekten er målt som Pearsons *r*. Merk at verdier av *d* og *r* ikke er sammenlignbare, utover at høyere verdier representerer sterkere effekt. Det finnes tommelfingerregler for tolkning av effektverdier som «sterke», «middels» eller «svake» tendenser (f.eks. Cohen, 1992b, s. 157), men generelt bør slike verdier tolkes i lys av kunnskap om studieobjektet (se f.eks. Field, Miles & Field, 2012, s. 58; Plonsky & Oswald, 2014; Wasserstein, Schirm & Lazar, 2019).

Tabell 1: Nødvendige utvalgsstørrelser for toutvalgstest og korrelasjonstest for ulike effektstørrelser oppgitt som henholdsvis Cohens *d* og Pearsons *r*.

Utvalgsstørrelser	**N=15**	**N=30**	**N=60**	**N=120**	**N=240**
Effekt (Cohens *d*)	1,06	0,74	0,52	0,36	0,27
Effekt (Pearsons *r*)	0,66	0,49	0,35	0,25	0,17

Note: Normalfordelte populasjoner, α=0,05, β=0,20, tosidige hypoteser. For toutvalgstest er N størrelsen av hvert av de to delutvalgene.

Tabellen viser at de utvalgene som er nødvendige for å avdekke effekter på $d \approx 0{,}27$ eller $r \approx 0{,}17$, er temmelig store – større enn i alle studiene nevnt ovenfor. I de fleste tilfeller er det dermed lite aktuelt med hypotesetestende studier for å avdekke så svake effekter, selv om vi skulle anta at de finnes og regne dem som faglig interessante. Konsekvensen av sammenhengene som tabellen viser mellom effekt og utvalgsstørrelse, er at man i hypotesetestende studier med utvalg på for eksempel mellom 30 og 60 informanter risikerer å forkaste hypoteser om effekter av faglig betydning utelukkende fordi utvalgene er for små.

Beregningene som gjelder toutvalgstester ovenfor, er basert på bruk av t-test på normalfordelte populasjoner, og beregningene med Pearsons korrelasjonstest er også på normalfordelte populasjoner. I de fire metaforstudiene er det andel riktige svar i flervalgsoppgavene som analyseres; dette er forholdstall, og forholdstall er gjerne ikke normalfordelte. Analyse av slike data forutsetter derfor ikke-parametriske tester (Jensen, 2018, s. 452–453). Ikke-parametriske tester som Wilcoxon-test og Spearmans korrelasjonstest har den ulempe at de når populasjonene er normalfordelte, har mindre styrke enn parametriske tester (Boneau, 1960; Chok, 2008), og det medfører at de trenger noe større utvalg for å oppnå samme styrke som en parametrisk test. Forskjellen i styrke er likevel ikke dramatisk; for balanserte utvalg på $N = 2x20$, $\alpha = 0{,}05$, $d = 1$ og normalfordelte populasjoner er styrken for en t-test $\pi \approx 0{,}87$, mens den for tilsvarende ikke-parametriske Wilcoxontest er $\pi \approx 0{,}85$.[9] For ikke-normale fordelinger er det mer komplisert å lage relevante sammenligninger, og måten fordelingene avviker fra

[9] Beregnet med 10 000 gjentatte simuleringer på tilfeldige utvalg trukket fra normalfordelte populasjoner. Beregningene ble utført av én av forfatterne (Jensen).

normalitet på, vil i seg selv påvirke styrken i beregningene. Det er imidlertid påvist at Wilcoxon-testen har mer styrke enn t-testen for visse typer fordelinger (Zimmerman & Zumbo, 1990), og tilsvarende gjelder for ikke-parametriske kontra parametriske korrelasjonstester (Bishara & Hittner, 2012). Det er derfor vanskeligere å beregne nødvendig utvalgsstørrelse dersom fordelingene avviker fra normalfordelingen, og særlig hvis avviket er stort.

Tallene for toutvalgstester i tabell 1 gjelder for to *balanserte* utvalg, altså utvalg som på grunnlag av en dikotom forklaringsvariabel er delt i to helt like store delutvalg. Dersom utvalget er skjevfordelt og ett av delutvalgene er mindre enn det andre, vil styrken reduseres. En sammenligning av de to delutvalgene av størrelse $N_1 = 61$ og $N_2 = 18$ i Golden og Larsen (2005) har altså mindre statistisk styrke enn den ville ha hatt om utvalget på $N = 79$ hadde vært delt i to nesten like store delutvalg $N_1 = 40$ og $N_2 = 39$ (men større styrke enn med $N_1 = 18$ og $N_2 = 18$). Dette er et viktig argument for å etterstrebe noenlunde balanserte delutvalg i datainnsamlingen dersom sammenligning av delutvalgene er blant forskningsspørsmålene, også om man bruker ett av utvalgene som et kontrollutvalg, f.eks. bestående av L1-talere, som i Golden og Larsen (2005).

I Golden (2005) dreier ett av forskningsspørsmålene seg om en sammenlikning av to typer av metaforiske uttrykk, for å finne om den ene typen er vanskeligere enn den andre å forstå, slik som enkeltordsuttrykk (verbet *snu* i uttrykket «*snu* i en sak») og flerordsuttrykk (som *slå alarm* i uttrykket «*slå alarm* om forurensning»). Andre sammenlikninger var mellom uttrykk som ble karakterisert som å ha ung og voksen tematikk, eller uttrykk med eller uten ekvivalenter i førstespråket. Det vil si at problemstillingene er grunnleggende *paret*, ved at de sammenlikner to og to observasjoner av de samme individene i motsetning til en sammenlikning av to grupper av individer. Til slike parede observasjoner brukes paret t-test eller paret Wilcoxon-test (eller fortegntest/binomialtest, se Golden (2010)). Det at testen sammenlikner verdier parvist, øker den statistiske styrken i testingen i forhold til tallene i tabell 1. Akkurat hvor stor økningen i styrke er, kommer an på de faktiske egenskapene til de parede dataene.

Komplekse forskningsspørsmål

Mange forskningsspørsmål forutsetter at man sammenlikner flere enn to kategorier. Statistiske analyser av data med tre eller flere kategorier i forklaringsvariabelen blir gjerne utført med enveis anova og en *post-hoc*-test. Anova har som nullhypotese at det ikke finnes forskjeller mellom noen av delutvalgene; den alternative hypotesen er dermed at det finnes en forskjell mellom minst to av dem, men uten at det er spesifisert hvilke to. Mer generelt tester en enveis anova ($k(k–1))/2$ hypoteser, der k er antall kategorier i forklaringsvariabelen. Som regel er ikke alle disse hypotesene relevante for studien – særlig ikke hvis k er stor. I så fall kan man øke styrken i beregningene ved å teste bare de spesifikke hypotesene man faktisk er interessert i. Jensen argumenterer på grunnlag av dette for å bruke anova først og fremst i eksplorativ tilnærming, og ikke som hypotesetest (2020, s. 54–55).[10] Et ikke-parametrisk alternativ til anova er Kruskal-Wallis' H-test (Jensen, 2018, s. 453–454).

Ofte har man flere enn én forklaringsvariabel, for eksempel L1 og ferdighetsnivå eller metaforer kategorisert i henhold til ulike kriterier, og man er interessert i potensielle interaksjoner mellom disse forklaringsvariablene. I likhet med når en forklaringsvariabel har flere verdier (se ovenfor), gir dette komplekse hypoteser som kan analyseres med to- eller flerveis anova, men som det kan være bedre å forsøke å redusere til færre og mer spesifikke hypoteser. Dessuten må man være oppmerksom på at den statistiske styrken reduseres dramatisk når man deler utvalget i stadig mindre delutvalg, jf. diskusjonen om utvalgsstørrelse ovenfor. Når delutvalgene er balansert, vil to interagerende dikotome forklaringsvariabler redusere størrelsen på delutvalgene til N/4; tre forklaringsvariabler gir N/8. I så fall gir et utvalg på totalt N = 120 delutvalg på bare $N_i = 15$; studier med så komplekse hypoteser blir altså i praksis svært krevende å gjennomføre.

En alternativ tilnærming for å redusere kompleksiteten i analysen kan være å holde noen av de potensielle forklaringsvariablene konstante, altså bruke dem som kontrollvariabler. Dette reduserer variasjonen i materialet og kan dermed gjøre effekten av de gjenværende forklaringsvariablene lettere å avdekke.

[10] Hypotesetestende statistiske tilnærminger bruker statistiske tester til å *bekrefte* eller *avkrefte* konkrete hypoteser om populasjoner, mens eksplorative tilnærminger bruker statistiske metoder og verktøy til å *utforske* og *øke forståelsen* av data som en fra før gjerne mangler tilstrekkelig kunnskap om til å utforme konkrete hypoteser. Se f.eks. Jensen (2020).

Utvelgelse av informanter

Studiene som er omtalt ovenfor, bruker informantutvalg av ganske ulike størrelser. Golden (2005, s. 81) har N = 170 og et kontrollutvalg på N_k = 230 med norsk som L1; Golden og Larsen (2005) har N = 61 og N_k = 18; Golden og Szymańska (2021) og Szymańska og Golden (u. a.) har henholdsvis N = 17 og N = 25 og ingen kontrollgruppe. Den statistiske styrken er dermed også temmelig ulik i disse studiene, og det lave antall informanter i kontrollutvalget hos Golden og Larsen er en faktor som reduserer styrken ytterligere. Golden og Larsen (2005) sammenlikner f.eks. elever som har gått i norsk barnehage (N_1 = 35), med elever som ikke har denne bakgrunnen (N_2 = 14), og finner en forskjell mellom gruppene som ikke er signifikant; i denne testen er styrken vesentlig svekket av det ene delutvalgets begrensede størrelse, og kanskje ville tendensen vist seg signifikant dersom dette delutvalget hadde vært større. I praksis kan det imidlertid være utfordrende å oppnå balanserte delutvalg, ettersom mange bakgrunnsopplysninger ikke vil være tilgjengelige før datainnsamlingen er gjennomført.

Golden og Szymańska (2021) og Szymańska og Golden (u. a.) reduserer kompleksiteten knyttet til antall variabler ved å benytte informanter med samme L1. Selv om studentene var både andre- og tredjeårsstudenter, viste det seg at det blant andreårsstudentene kun var dem med de beste norskferdigheter som valgte å levere eller å fylle ut hele testen, og ferdighetsnivået deres i norsk ble relativt likt. I praksis kunne derfor både førstespråket og ferdighetsnivået fungere som kontrollvariabler. Dette opprettholder utvalgsstørrelsene og reduserer variasjonen i materialet; dermed blir de gjenværende forklaringsvariablene potensielt mer fremtredende i variansbaserte analyser som t-test og anova.

Utformingen av oppgavesettet

Det første skrittet i utformingen av et oppgavesett er å velge testuttrykk som er egnet til å belyse forskningsspørsmålet. Hvis forskningsspørsmålet er knyttet til forståelse av ord og uttrykk fra et visst temaområde eller en sjanger, vil dette være styrende for utvalget. I Golden (2005) var forskningsspørsmålet knyttet til metaforiske ord og uttrykk som fore-

kom i skolebøker og var nødvendige for forståelsen av slike læreboktekster. Derfor ble uttrykkene hentet fra samfunnskunnskapsbøker. Uttrykk fra to begrepsmetaforer ble bevisst tatt med, fordi de viste seg å være frekvente i bøkene, men ellers var det ikke fokus på noen spesielle typer. Uttrykkene var varierte siden det var ønskelig «å vise mangfoldet av uttrykk som fantes i lærebøkene» (Golden, 2005, s. 86), og det ble ikke tatt hensyn til at uttrykkene var svært forskjellig oppbygd eller hvor bildesterke (gjennomsiktige) de var. Dette førte til at analysen ble kompleks (se nedenfor).

Kontekst

Når den overordnede hensikten er å studere leseforståelse, er det viktig å oppgi ordene i en kontekst, siden ordforståelse er kontekstavhengig. Dette gjøres vanligvis ved en innledning, som kan være av ulike lengder. Selv om Golden hentet uttrykkene fra skolebøker, framhever hun at det ikke var mulig å bruke de autentiske setningene som de metaforiske uttrykkene var del av, ettersom det ville gi for vag kontekst (Golden, 2005, s. 91). Da måtte elevene få presentert en lengre del av teksten, noe som ville ha økt lesetiden for hver oppgave og dermed gjort det nødvendig å redusere antall oppgaver for å holde seg innenfor ønsket tid. Ortografisk usikkerhet spiller ofte en rolle hos innlærere. En innledning – om den er kort (en del av en setning) eller lang (flere setninger) – kan bøte på dette ved at den frembringer en situasjon som gjør testuttrykket entydig og reduserer risikoen for forveksling av homografer eller formlike ord. Imidlertid kan innledningen skape nye utfordringer; den kan inneholde nye og kanskje ukjente ord, og den kan inneholde ord som gir feil assosiasjoner. Med andre ord har man mindre og mindre kontroll på forståelsen av konteksten etter hvert som den blir lengre, noe som gjør tolkningen av resultatene mer utfordrende. Strukturen i innledningen vil også variere, og den kan karakteriseres som kompleks eller enkel avhengig av hvor lett den er å prosessere, og den kan presentere et tema som passer for noen aldersgrupper, men ikke andre. Ett av resultatene i Golden (2005) var nettopp at det var statistiske forskjeller i forståelsen av enkeltoppgavene med såkalt ungt og voksent tema. Til tross for at utformingen av de enkelte innledningene var ment å ta hensyn til elevenes alder, virket det som at deres interesser, som Golden antok særlig dreide seg om relasjoner i skolehverdagen, påvirket forståelsen.

Selv om kontekst er ønskelig når ordforrådet testes som ledd i leseforståelse, er det ikke uinteressant å undersøke forståelsen av uttrykk uten kontekst. Det er riktignok sjelden at ord forekommer alene, men det skjer i titler og på plakater. I skolebøker har titler også en pedagogisk hensikt, ettersom de sier noe om hva teksten handler om. I en tekst Kulbrandstad (1996) brukte i sin undersøkelse av leseforståelse, var tittelen det metaforiske uttrykket *Medaljens bakside*. Forståelsen av dette uttrykket ville gi elevene en inngang til tekstens innhold, og slik sett kan uttrykk uten kontekst også være interessant å undersøke.

Distraktorer

I flervalgstester varierer antall distraktorer, og selv om flere distraktorer kan gjøre oppgaven vanskeligere, vil ordvalget i de enkelte distraktorene spille den største rollen. Rodriguez (2005) har gjennomgått en rekke flervalgsstudier og mener at man bør lage så mange distraktorer som mulig, men hevder likevel at det har vist seg vanskelig å lage flere enn tre gode. En gyllen regel er at distraktorene ikke må være vanskeligere enn testuttrykket. Ifølge Nation (2001) avhenger vanskelighetsgraden til oppgaven av hvor nær distraktorene ligger det riktige svaret i betydning. De må altså ses både i forhold til testuttrykket, til hverandre og i tillegg til innlærernes språklige repertoar. I materialet som ligger til grunn for de fire eksempelstudiene, er det tilstrebet, men ikke gjennomført, at én av distraktorene skulle inneholde et ord med en *form* som lignet testordet. I Golden (2005) ble noen av feilene regnet for å være resultat av en formkoblingsstrategi: det virket som om noen elever sammenliknet ordformen i oppgavestammen og svaralternativene og valgte det svaralternativet hvor de fant den største formlikheten uavhengig av innhold. I en oppgave stod det f.eks. «å synes vennskap er *verdifullt*» i oppgavestammen, og mange krysset av for alternativet «liker å kjøpe verdifulle ting». I en annen lød oppgavestammen «*komme* deg *i møte* i en diskusjon», og her valgte mange «dere skal møtes etterpå». Slike svar kan tyde på at elevene har gjettet.

Distraktorenes avstand til hverandre og til det rette alternativet er avhengig av hensikten, hvorvidt man studerer *bredde* eller *dybde* i ordforrådet. Bredde viser til kvantitet, dvs. antall ord en kjenner til eller mestrer, mens dybde kan konseptualiseres som en glidende skala fra *ikke-kjennskap* til *fullstendig mestring* av ord og uttrykk, og ulike forskere har gradert denne skalaen forskjellig (jf. Paribakht & Wesche, 1997;

Wesche & Paribakht, 1996; Read, 2000, 2004; Schmitt & Zimmerman, 2002). Svaralternativenes semantiske avstand vil avgjøre hvor stor dybde en tester; jo mindre avstand det er mellom dem, jo større dybde i forståelsen av ordet kreves for å velge riktig alternativ. I oppgavesettet som ligger til grunn for de fire studiene, er det varierende dybde som testes ved at avstanden mellom det semantiske innholdet i svaralternativene varierer. For eksempel forekommer «dårlig argument» og «innholdsløst argument» i tillegg til «viktig argument» og «nytt argument» som forklaring på hva et «*tungt* argument» betyr. Her er to av forslagene negative («dårlig» og «innholdsløst»), ett er positivt («viktig») og ett nøytralt («nytt») når det gjelder å vise hva vurderingen av et argument som «*tungt*» vil si. Avstanden mellom de to negative forklaringene er altså mindre enn mellom de andre forslagene, og hvis ett av dem hadde vært riktig, ville det ha vist større dybde i forståelsen å velge det ene. Det samme gjelder et annet uttrykk som testes, «*glødende* interessert»: det er større avstand mellom de to forslagene «litt interessert» og «veldig interessert» enn mellom de andre «uinteressert» og «likegyldig». Oppgaven med uttrykket «elevene *brenner* for noe» har derimot stor avstand mellom alle forslagene til forklaring: at de «er veldig interessert i noe», «lager bål», «lager hull i noe» eller «er veldig opptatt av brann». Her trengs ikke like stor dybde i forståelsen for å velge rett alternativ. Dette bør man være bevisst i analysen. Rekkefølgen på alternativene spiller også en rolle; ifølge Cohen, Manion og Morrison (2011) foretrekkes det første forslaget hvis man er usikker.

Parallelle uttrykk i L1 og L2

Begrepsmetaforer som i utgangspunktet hevdes å være universelle, kan bli realisert av et metaforisk uttrykk som ikke blir forstått fordi det benytter andre leksemer enn i det tilsvarende uttrykket i førstespråket eller at noen av leksemene er ukjente. En kartlegging av ekvivalens mellom innlærerens L1 og L2 er derfor nødvendig hvis en vil finne ut hvilken betydning dette har for forståelsen. I Golden (2005) ble det undersøkt om det fantes uttrykk i språkene punjabi/urdu, tyrkisk og vietnamesisk som tilsvarte det norske uttrykket. Dette ble gjort ved en dobbel oversettelsesmetode: en person oversatte fra norsk til sitt førstespråk, og en

annen person oversatte dette uttrykket tilbake til norsk.[11] Avhengig av hvor like disse uttrykkene var (både når det gjaldt innhold eller ordbruk), ble uttrykkene regnet for å være ekvivalente eller ikke ekvivalente. Dette skillet mellom likhet og ulikhet er altså binært hos Golden (2005). Men både i Golden og Szymańska (2021) og i Szymańska og Golden (u. a.) var formålet å undersøke hvorvidt ulike grader av likhet påvirket forståelsen, og særlig om delvis likhet var en støttende eller forstyrrende faktor. Derfor ble det foretatt en tredeling. Kategoriene var: uttrykk *med fullstendig ekvivalens* i polsk (f.eks. *å tråkke på hverandre = deptać po sobie* / 'tråkke på hverandre-INST'), uttrykk *med delvis ekvivalens* i polsk og uttrykk *uten ekvivalens* i polsk (f.eks. *å føre noen bak lyset*). I delvis-kategorien ble det plassert uttrykk som liknet på de norske uttrykkene enten ved valg av ord, som i *å trekke en ned i søla = zmieszać kogoś z błotem* / 'å blande noen-AKK med søle-INST' eller ved at det viste til det samme bildet, f.eks. *å vrake kjæresten = rzucić chłopaka/dziewczynę* / 'kaste gutt/jente-AKK'. Det er imidlertid vanskelig å bestemme grensene for uttrykk i denne kategorien, særlig når det gjelder flerordsuttrykk. Dersom man legger vekt på strukturelle trekk, som rekkefølgen av de leksikalske enhetene, er det veldig få uttrykk som er identiske i begge språkene, siden de grammatiske forholdene er ulike. Enkelte leksemer kan f.eks. få ulik plassering (slik som *medaljens bakside* og *druga strona medalu* / 'andre side medalje-GEN', eller *å stole blindt på noen* og *ślepo komuś ufać* / 'blindt noen-DATIV stole'), som igjen kan gjøre at innlæreren bare fokuserer på en del av uttrykket og ikke på hele. Da kan åpenbare likheter, f.eks. på substantivnivå, bli oversett. I en studie av forståelse av engelske metaforiske uttrykk hos informanter med ulike L1 skiller Zimmermann (2006) mellom fire typer ekvivalens av metaforiske uttrykk: 1: uttrykk med samme *mapping*, dvs. overføring mellom de samme to domenene med bruk av tilsvarende leksikalske elementer, 2: uttrykk med samme mapping, men med ulike leksikalske elementer, 3: uttrykk uten ekvivalens i innlærerens førstespråk, og 4: grensetilfeller. Kategoriseringen likner på den hos Golden og Szymańska (2021) og Szymańska og Golden (u. a.), men Zimmermann skiller mellom uttrykk med ulike leksikalske elementer og det han kaller grensetilfeller, mens Golden og Szymańska (2021) og Szymańska og Golden (u. a.) plasserer delvis ekvivalente uttrykk i én og samme kategori, uavhengig av hva forskjellen består i. Imidlertid peker også

[11] Der usikkerheten var for stor, ble ikke uttrykket kategorisert.

Zimmermann på at det er vanskelig å bestemme hvor grensen mellom de ulike kategoriene går (se Golden og Szymańska (2021) for ytterligere diskusjon).

Dersom man velger å se bort fra det strukturelle i ekvivalensdiskusjonen og antar at det er de leksikalske elementene som har mest å si for skillet mellom fullstendig og delvis ekvivalens, er det innholdsordene, dvs. substantiv, verb og adjektiv, som er avgjørende. Det kan diskuteres hvilken ordklasse av de tre som påvirker uttrykkets gjenkjennelighet i størst grad, men Zimmermann (2006) regner substantivet som det leksikalske kjerneelement. Hvis dette kjerneelementet er ulikt i de to språkene, kan det gi betydningsnyanser. Et godt eksempel finner vi i Horbowicz (2005, 2009), som undersøkte norske og polske metaforer med kropp som kildedomene. Hun trekker fram uttrykket *hva har du på hjertet? = co ci leży na wątrobie?* / 'hva deg-DATIV ligger på lever-LOC', hvor *lever* i det polske uttrykket gir negative konnotasjoner, mens *hjerte* i det norske uttrykket er nøytralt/positivt. I utgangspunktet er det to uttrykk med noenlunde tilsvarende semantisk innhold som brukes i tilnærmet identisk sammenheng, men bruken av et annet substantiv kan føre til at betydningen av uttrykket blir vanskelig å gjennomskue for en polsk innlærer. På norsk har man uttrykkene *is i magen* eller *holde hodet kaldt*, mens på polsk er det *blod* som må holdes *kaldt*: *zachować zimną krew* / 'beholde kald blod-AKK'. Man *skyter seg selv i foten*, mens på polsk *skyter man seg i kneet – strzelić sobie w kolano* / 'skyte seg selv i kne-AKK'. Hvorvidt bruk av ulike leksikalske kjerneelementer er viktig for forståelsen, er uvisst, men dette illustrerer hvor vanskelig det er å skille mellom fullstendig og delvis ekvivalente uttrykk.

I tillegg til uttrykk hvor variasjonen holder seg innenfor det samme domenet, finnes det en rekke uttrykk som har identisk betydning i begge språkene, men bruker ord fra helt forskjellige domener, som *slå to fluer i én smekk = upiec dwie pieczenie na jednym ogniu* / 'steke to stykker kjøtt-AKK på en ild-INST', eller bruker andre karakteristikker av samme ordet som *skarpe hoder = tęgie głowy* / 'fete el. sterke hoder', eller bruker andre liknende, men ikke identiske uttrykksmåter, som *å gå hånd i hånd = iść w parze* / 'gå i par-LOC'. Uttrykk som har samme innhold i polsk, til tross for bruk av ulikt leksem, ble i Golden og Szymańska (2021) og Szymańska og Golden (u. a.) også klassifisert som delvis ekvivalente. Denne diskusjonen viser hvor viktig, men vanskelig

det er å etablere gode kriterier som kan brukes konsekvent ved kategorisering av ekvivalens.

Analyse

For å finne ut om studentene kjente til betydningen eller kun gjettet svaret, ble de i Szymańska og Golden (u. a.) bedt om å krysse av for dette etter hvert svar. Resultatet her var uventet: det var flere riktige svar hos dem som oppgav å ha gjettet, enn hos dem som ikke hadde gjort det. Dessuten var det store forskjeller mellom studentene; noen hadde krysset av for gjetting svært mange ganger, mens andre aldri oppgav å ha gjettet. Szymańska og Golden tolket dette som forskjell i studentenes teststrategier eller i innlærernes vilje eller mot til å innrømme usikkerhet. Dette viser at selvrapportering ikke er hensiktsmessig i slike studier. Et alternativ kan være å legge til «vet ikke» som et svaralternativ, men muligens vil dette gi samme resultat. Prentice (2010) anbefaler i tråd med Nagy, Herman og Anderson (1985) å konstruere flervalgstester med ulik vanskelighetsgrad, ved at den semantiske avstanden mellom distraktorene og det rette svaret varierer. Slik sett vil en del av oppgavene være enkle, og da kan det være lettere å innrømme usikkerhet på de vanskelige oppgavene. Imidlertid hevder Jones (2021) at vill gjetting sjelden forekommer; hvis deltakerne har nok tid, er det alltid noe i et alternativ som påkaller oppmerksomheten. Dessuten ville en slik bevisst variasjon i vanskegrad komplisere tolkningen av resultatene.

De fire studiene involverer forklaringsvariabler som består av flere enn to kategorier. Szymańska og Golden (u. a.) undersøker for eksempel hvorvidt forståelsen varierer avhengig av om det metaforiske uttrykket har fullstendig, delvis eller er uten ekvivalens i polsk, mens Golden (2005) og Golden og Larsen (2005) sammenlikner flere grupper av informanter, basert på selvrapporterte bakgrunnsopplysninger. Slike forskningsspørsmål med flere kategorier kan som nevnt angripes med anova, med de ulemper dette medfører. I Szymańska og Golden (u. a.) er imidlertid hypotesen mer spesifikk: forståelsen av metaforene i fullstendig ekvivalens-kategorien antas å være bedre enn i delvis-kategorien, og forståelsen av disse bedre enn i uten ekvivalens-kategorien. Gjennom å teste disse to delhypotesene hver for seg kunne styrken i testingen økes ved å benytte to parede toutvalgstester i stedet for anova eller Kruskal-

Wallis' H-test. Når komplekse forskningsspørsmål kan brytes ned til noen få parvise sammenligninger på denne måten, øker det styrken i analysene.

Golden (2005) undersøkte elever med ulike L1, og hun delte dessuten elevene inn i tre kompetansegrupper etter hvilket språk de selv mente de kunne best: a) de som mente de kunne norsk best, b) de som mente de kunne morsmålet best, og c) de som mente de kunne begge språkene like godt. Golden testet imidlertid ikke disse faktorene i én modell. Hun sammenlignet derimot L1-gruppene og kompetansegruppene hver for seg, uten eksplisitt å teste interaksjonshypoteser, altså for eksempel en hypotese om at forskjeller mellom visse L1-grupper er sterkest for visse kompetansegrupper. Hun unngikk dermed utfordringene med små delutvalg som kan oppstå ved flere samtidige forklaringsvariabler, nevnt i teoriseksjonen ovenfor.

Som forklart tidligere er det alltid en risiko for at en negativ konklusjon fra en hypotesetest skyldes for liten statistisk styrke i studien, og mangel på statistisk styrke kan ofte skyldes at utvalget er for lite for formålet med studien, enten det skriver seg fra en kompleks hypotese eller et høyt antall hypoteser. Negative resultat på bakgrunn av lav statistisk styrke er uheldige fordi det lett kan oppfattes som en reell tilbakevisning av hypotesen. Man bør derfor alltid gjøre beregninger av statistisk styrke under planleggingen av studien før man gjennomfører hypotesetesting. Dersom styrkeberegningene viser at studien ikke kan avdekke de effektstørrelser man er interessert i, bør man heller avstå fra å utføre hypotesetesting og følgelig heller ikke rapportere p-verdier eller signifikans. I stedet bør man rapportere effektstørrelser på en fornuftig måte: For sammenlikning av utvalg bør det rapporteres både nominelle differanser og et relevant effektmål som Cohens *d*. Både differanser og effekter skal følges av konfidensintervaller (CI); for alle konfidensintervaller er 95 % sikkerhet et vanlig og fornuftig valg. For eksempel oppgir Golden og Szymańska (2021) den nominelle differansen (D) mellom andel forståtte ettordsuttrykk og andel forståtte flerordsuttrykk til $D \approx 0{,}120$, med konfidensintervall [0,082, 0,16]. Det betyr at differansen i utvalget er 0,120, mens konfidensintervallet kan forstås slik at man kan være 95 % sikker på at den reelle differansen i populasjonen ligger et sted mellom 0,082 og 0,16.[12] Den tilsvarende effekten er

[12] Det er praktisk å tenke på konfidensintervaller på denne måten, selv om formuleringen ikke er helt matematisk presis.

rapportert som $d \approx 0{,}98$, med konfidensintervall [0,62, 1,34]. For korrelasjoner er korrelasjonskoeffisienter som Pearsons *r* eller Spearmans *rho* egnet som effektmål; også korrelasjonskoeffisienter skal rapporteres med konfidensintervall. Korrelasjonen mellom andel forståtte konkrete uttrykk og andel forståtte metaforiske uttrykk ble av Golden og Szymańska (2021) oppgitt til $r \approx 0{,}57$, med 95 % konfidensintervall på [0,13, 0,82]. Konfidensintervallet viser svært stor usikkerhet når det gjelder styrken i den rapporterte korrelasjonen, fra ubetydelig til svært sterk.

Avslutning

Flervalgstest er generelt en mye brukt metode, og dens fordeler er særlig knyttet til kostnad. Ikke bare kan mange informanter testes relativt raskt og uten store ressurser, men testen kan også enkelt omfatte mange enkeltoppgaver, noe som er en stor fordel nettopp ved testing av et fenomen så mangfoldig og omfattende som forståelse av ord og uttrykk, og spesielt når de er metaforiske. Til tross for dette er det ikke mange studier av innlæreres forståelse av norsk ordforråd som har brukt flervalgstester. Kanskje skyldes dette at analysene lett kan bli komplekse; særlig metaforiske uttrykk varierer på mange måter, f.eks. i lengde, gjennomsiktighet og syntaktisk struktur, og alle slike variabler kan påvirke forståelsen. Dersom alle forklaringsvariabler skal inkluderes i analysen, reduserer det den statistiske styrken i beregningene, og det kan forutsette innsamling av data fra ganske store utvalg av informanter.

Ordforrådskunnskap er et komplekst konstrukt, og for å kunne vurdere de ulike aspektene ved denne kunnskapen mer inngående trengs åpenbart både flere måleredskaper og flere typer mål (Schmitt, 2010, s. 152). Flervalgstester av metaforforståelse bidrar med én brikke i dette puslespillet, men grundig planlegging og bevisst utforming av oppgavesettet er nødvendig for at denne brikken skal være til hjelp i helhetsbildet.

Referanser

Ash, G. E. & Baumann, J. F. (2017). Vocabulary and comprehension: The nexus of meaning. I S. E. Israel (Red.), *The Handbook of Research on Reading Comprehension* (2. utg., s. 141–155). New York: Guilford.

Askeland, N. (2019). *Metaforer: Hva, hvor og hvorfor?* Oslo: Universitetsforlaget.

Bishara, A. J. & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *American Psychological Association*, *17*(3), 399–417. doi:10.1037/a0028087.

Boneau, A. C. (1960). The effects of violations of assumptions underlying the t-test. *Psychological Bulletin*, *57*(1), 49–64.

Chok, N. S. (2008). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data* (Masteroppgave). University of Pittsburgh.

Cohen, J. (1992a). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98–101.

Cohen, J. (1992b). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, L., Manion, L. & Morrison, K. (2011). *Research methods in education* (7. utg.). New York: Routledge.

Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. London: Sage.

Golden, A. (2005). *Å gripe poenget: Forståelse av metaforiske uttrykk fra lærebøker i samfunnskunnskap hos minoritetselever i ungdomsskolen*). Oslo: UniPub. https://www.hf.uio.no/multiling/personer/kjernegruppe/agolden/golden_hf_til_trykk.pdf.

Golden, A. (2010). Jeg fant, jeg fant – men hva gjør jeg med det? I H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (Red.), *Systematisk, variert, men ikke tilfeldig: Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag* (s. 113–119). Oslo: Novus.

Golden, A. (2014). *Ordforråd, Ordbruk, Ordlæring*. Oslo: Gyldendal Akademisk.

Golden, A. & Larsen, V. (2005). Egenvurdering av språkferdigheter og metaforforståelse blant minoritetselever i videregående skole. *NOA – Norsk som andrespråk*, *21*(1–2), 67–91.

Golden, A. & Szymańska, O. (2021). Griper polske studenter poenget på norsk? Forståelse av metaforiske uttrykk blant norskstuderende i Polen. Sendt *Nordand – Nordisk tidsskrift for andrespråksforskning*, under vurdering.

Horbowicz, P. (2005). *Kroppsdeler i følelsers tjeneste – en kognitiv analyse av faste uttrykk med kroppsdeler i norsk og polsk* (Masteroppgave). Adam Mickiewicz Universitet, Poznań.

Horbowicz, P. (2009). Varmt hjerte og kaldt blod. En kontrastiv analyse av uttrykk for følelser basert på indre organer og kroppsvæsker i polsk og norsk. *Folia Scandinavica Posnaniensia*, *10*, 169–180.

Hughes, A. (2003). *Testing for language teachers* (2. utg.). Cambridge: Cambridge University Press.

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

Jensen, B. U. (2020). Hypotesetesting versus utforskende statistikk i språkforskning. *NOA – Norsk som andrespråk*, *36*(1), 39–71.

Jones, G. (2021). Designing Multiple-Choice Test Items. I P. Winke & T. Brunfaut (Red.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (s. 90–101). New York: Routledge.

Kulbrandstad, L. I. (2018). *Lesing i utvikling. Teoretiske og didaktiske perspektiver (2 utg.)*. Bergen: Fagbokforlaget.

Lakoff, G. (1993). The contemporary theory of metaphor. I A. Ortony (Red.), *Metaphor and thought* (s. 202–251). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139173865.013.

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

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

Nagy, W. E., Herman, P. A. & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233–253.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Paribakht, T. S. & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. I J. Coady & T. Huckin (Red.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (s. 174–200). Cambridge: Cambridge University Press.

Plonsky, L. & Oswald, F. L. (2014). How Big Is «Big»? Interpreting Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079.

Prentice, J. (2010). *Käppen i hjulen: Behärskning av svenska konventionaliserade uttryck bland gymnasieelever med varierande språklig bakgrund.* Rapporter om svenska som andraspråk (ROSA) 12. Göteborg: Göteborgs Universitet.

Pripp, A. H. (2017). Antalls- og styrkeberegninger i medisinske studier. *Tidsskriftet den norske legeforening*, *137*(17), 1326. https://doi.org/10.4045/tidsskr.17.0414.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How Should the Construct of Vocabulary Knowledge Be Defined? I P. Bogaards & B. Laufer (Red.), *Vocabulary in a Second Language* (s. 209–227). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.10.15rea.

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. New York: Palgrave Macmillan.

Simmons, A. D., Rupley, W., Simmons, D. & Graham, L. (2011). The State of Vocabulary Research. *Literacy Research and Instruction*, *50*(4), 253–271.

Stevenson, M. (2010). Researching Reading. I B. Paltridge & A. Phakiti (Red.), *Continuum Companion to Research Methods in Applied Linguistics* (s. 174–190). London: Continuum.

Sweet, A. P. & Snow, C. E. (Red.) (2003). *Rethinking Reading Comprehension*. New York: The Guilford Press.

Szymańska, O. & Golden, A. (u. a.). Er *å holde hodet kaldt* det samme som *å beholde blodet kaldt?* Metaforiske uttrykk i norsk og polsk – Forståelse og ekvivalens.

Utdanningsdirektoratet (2011). *Det felles europeiske rammeverket for språk: Læring, undervisning, vurdering*. Oslo: Utdanningsdirektoratet.

Vosniadou, S., Ortony, A., Reynolds, R. E. & Wilson, P. T. (1984). Sources of difficulty in the young child's understanding of metaphorical language. *Child Development*, *55*(4), 1588–1606. https://doi.org/10.2307/1130028.

Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. (2019). Moving to a world beyond «p < 0.05». *The American Statistician*, *73*(sup1.), 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Wesche, M. & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge: Depth versus Breadth. *Canadian Modern Language Review*, *53*(1), 13–39.

Schmitt, N. & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*(2), 145–171.

Zimmerman, D. W. & Zumbo, B. D. (1990). The relative power of the Wilcoxon-Mann-Whitney test and Student t test under simple bounded transformations. *Journal of General Psychology*, *117*(4), 425–436.

Zimmermann, R. (2006). Metaphorical Transferability. I J. Arabski (Red.), *Cross-linguistic Influences in the Second Language Lexicon* (s. 193–209). Bristol: Multilingual Matters.

Abstract

In this article, we discuss various aspects of using multiple choice tests as a method in vocabulary studies, focusing on learners of Norwegian. To illustrate the discussion, four studies in which this method has been applied to investigate comprehension of Norwegian metaphorical expressions are presented, and different methodological issues related to these studies are discussed. The discussion includes considerations on statistical analysis in this kind of studies and potential challenges that may result from methodological choices taken. The article focuses mainly on selecting expressions to be tested, constructing appropriate distractors for the test items, the impact of the context, and statistical power. As our focus is on learners, we also discuss the concept of equivalence in metaphorical expressions between languages, as well as im-

plications of such equivalence for testing and analysing test results. The discussion on equivalence is illustrated by comparisons of metaphoric expressions in Norwegian and Polish.

Keywords: *multiple choice tests, metaphor, metaphoric expression, statistical power, Norwegian as a second language, Norwegian as a foreign language, cross-linguistic equivalence*