Høgskolen i Hedmark

Avdeling for lærerutdanning og naturvitenskap

Pål-Otto Mikkelsen

Masteroppgave

Digital sjibbolett? Testing av andrespråksskriving med penn og PC

Digital Shibboleth? Testing Second Language Writing Using Pen and PC

Master i kultur- og språkfagenes didaktikk

2016

Samtykker til utlån hos høgskolebiblioteket JA ☒ NEI ☐

Samtykker til tilgjengeliggjøring i digitalt arkiv Brage JA ☒ NEI ☐

Forord

Mye spennende forskning skjer i skjæringspunktet mellom vitenskaper. Denne avhandlingen er også et resultat av det. Fordi jeg fikk ta faget «Språk i digitale sjangre» har oppgaven blitt slik den er, så takk til studieledere i MIKS og Digkom. Mine veiledere, Marte Monsen og Bård Uri Jensen, skal også ha stor takk. Bård fordi han er nøyaktig, detaljfokusert og inspirerende, og Marte fordi hun er pragmatisk, forståelsesfull og ser litt stort på det, slik at jeg skjønner at det er håp. Dere utfyller hverandre godt, takk for det!

Takk til Inger Lise Stieng, Gunhild Tveit Randen, Lars Anders Kulbrandstad og Anne Golden som ga meg gode innspill på oppgaven i en tidlig fase. Lars Anders har også ellers vært entusiastisk, midt under en stor forskningskonferanse som han hadde ansvaret for, ville han ta en prat om min oppgave!

Biblioteket ved Campus Hamar og i Elverum skal ha takk for arbeidet med å skaffe bøker og artikler til arbeidet. Rektor, lærerne og elevene som deltok i prosjektet har vært tålmodige og samarbeidet godt, de fortjener en takk. Mine kolleger Lena og Linda har støttet med praktisk hjelp, lån av bøker og nyttige samtaler, takk for det!

Mine foreldre har inspirert meg til å være interessert i alt, og særlig språk. De har vært viktige diskusjonspartnere også her, og fortjener en stor takk. Også min bror fortjener en takk for teknisk assistanse. Den største takken går til familien, Håkon, Erik og Gøril, fordi jeg har fått sitte i flere ferier og mange helger og kvelder med oppgaven. Takk!

Eventuelle mangler og feil får stå for min regning.

Det skulle vært unødvendig å kjempe. Man skulle bare fått puste på sitt eget språk.

- Mari Boine

Innhold

Norsk sammendrag

Delferdigheten skriftlig produksjon på Norskprøven for innvandrere, arrangert av Vox, ble digitalisert i 2014. Antall elever som nådde nivå B1 (fra Det felles europeiske rammeverket for språk) i denne delferdigheten sank betydelig. Dette gjorde at jeg ville undersøke om det var kravet om digital skriving som forårsaket dette fallet. Jeg anvendte ti måleenheter på syntaks og ordforråd i språket og analyserte dem kvantitativt. Måleenhetene er hentet fra forskningstradisjonen som ser på kompleksitet, nøyaktighet og flyt som kvalitetskriterier i språket. Med bakgrunn i et sosialt syn på testing anvendte jeg teorier hentet fra Messick (1989) og Bachman (1990) og så på konstruktvalideringen av Norskprøven for å vurdere om kravet om digital skriving forfordeler elever som ikke har tilstrekkelige ferdigheter i å skrive på PC, med et særlig blikk på eldre voksne.

Teorier om digital skriving sier at elevene får mer tid til planlegging og revidering når de ikke må bruke så mye tid på transkribering og at de får dermed lengre og bedre tekster. Jeg fant svært små forskjeller og større ulemper enn fordeler for de eldste deltakerne. Dette kan skyldes at elevene ikke klarer å utnytte fordelene med digital skriving. Andrespråksskriving er allerede så krevende at mye kognitiv kapasitet går med til formulering og det blir lite tid til planlegging og revidering. Elever som har høy skrivehastighet på PC synes i noen grad å utnytte denne til økt lengde, men særlig til å redusere antall feil. Noen funn spriker i begge retninger, og kan indikere at disse målene ikke er anvendelige i slike undersøkelser. Antall deltakere var også lite, det kan også være årsaken til at resultatene ikke ble signifikante. Mulige konsekvenser kan være at Vox informerer deltakere og lærere tydeligere om kravet til digital skriving, og at lærere ved voksenopplæringer bruker mer tid på skriveprosessen og på eksplisitt undervisning i digital skriving. Det bør også vurderes utvidet skrivetid på Norskprøven, bruk av retteprogram, eller om enkelte kandidater kan gjennomføre prøven med penn og papir i tråd med prinsippene for rettferdig testing (Messick, 1998). En endring av konstruktet eller Norskprøven kan også være mulige konsekvenser.

Engelsk sammendrag (abstract)

The subskill written production in the language test for immigrants, Norskprøven for innvandrere, arranged by Vox, was digitized in 2014. The number of students reaching the B1 level (from the Common European Framework of Reference for Languages) in this subskill declined considerably. This moved me to explore whether the demand for digital writing caused the decline. Ten measures of syntax and lexicon from the scientific tradition regarding complexity, accuracy and fluency as criteria of quality in language were employed, these were analyzed quantitatively. A social view of language testing with theories from Messick (1989) and Bachman (1990) was applied. I looked at the construct validation of the language test Norskprøven for innvandrere to consider whether the test treats pupils with insufficient computer writing skills unfairly, especially regarding older adults.

Digital writing theories posit an increase in the time allotted for students' planning and revising due to a reduction in the time spent transcribing, hence increasing length and quality of their texts. I revealed small differences in the results, with disadvantages outweighing the advantages for older participants. This may be caused by students not being able to utilize the advantages of digital writing. Second language writing is demanding by itself, and students' cognitive capacity may already be preoccupied by formulation, leaving too little time for planning and revising. Students with a high writing speed on computer seem to increase text length to a certain extent, but more markedly to reduce errors. Some findings are inconclusive, possibly indicating that the measures are less applicable in such inquiries as the present. The number of participants also was small, thus significant results were not gained. Possible consequences may be that Vox inform students and teachers more explicitly on the demands of digital writing, and that teachers leave more time for the writing process and explicit teaching in digital writing. One should also consider an expansion in writing time on Norskprøven, the use of a spellchecker, or whether some candidates might use pen and paper in the test following the principles for fair testing (Messick, 1998). A change in the construct or Norskprøven may also be possible consequences.

Tabelliste

1. Innledning

Innvandrere til Norge som ønsker permanent oppholdstillatelse eller statsborgerskap i landet, må delta i norsk- og samfunnsfagopplæring. Også personer som er arbeidsinnvandrere kan delta i dette selv om de ikke ønsker permanent opphold eller statsborgerskap. De fleste avslutter denne opplæringen med en norskprøve. I denne oppgaven vil jeg se på Norskprøven slik den er i dag, og jeg vil bruke begrepene *prøve* og *test* synonymt. Jeg vil se om skriveredskapet har en betydning for hvilket nivå kandidatene oppnår i delferdigheten skriftlig framstilling. Det er avsluttende norskprøver på flere nivåer for å dokumentere oppnådd nivå, enten for opptak på en skole eller for en arbeidsgiver. For personer som har fått rett og plikt til opplæring etter 1. september 2013 er det obligatorisk å delta på norskprøve. Sammen med økende innvandring og at prøvene har blitt mer kjent betyr dette at antall kandidater til norskprøvene har økt de siste årene. Norskprøvene består av fire delprøver som måler ulike ferdigheter: *muntlig* (som igjen består av *å snakke* og *å samtale*), *lytteforståelse, leseforståelse* og *skriftlig framstilling*. I 2015 ble det avholdt over 55 700 slike delprøver, mer enn 15 000 av disse var i delferdigheten *skriftlig framstilling* (Vox, 2016d). Vox er et nasjonalt fagorgan for kompetansepolitikk underlagt Kunnskapsdepartementet, og er faglig ansvarlig for voksenopplæring og norsk- og samfunnsfagprøvene.

Med et samfunn som i økende grad er teknologisk, var det naturlig at Vox innlemmet digital kompetanse som en basiskompetanse i den reviderte læreplanen i norsk og samfunnskunnskap for voksne innvandrere (Vox, 2012). Fra mai 2014 har også Norskprøven blitt endret, slik at kandidatene må bruke PC på den ene delen (leseforståelse, lytteforståelse og skriftlig produksjon). Den muntlige delen er uforandret. Jeg vil se på det som nå heter Språkprøven A2/B1, som er en videreføring og utvidelse av Norskprøve 2 og Norskprøve 3. Uttrykket *Norskprøven* er også fortsatt i bruk. Da de nye prøvene ble innført, økte antall kandidater som ikke nådde B1-nivå i skriftlig produksjon. Omtrent 50 % besto Norskprøve 3 skriftlig tidligere, men etter at den nye prøven kom, sank antall elever som nådde B1-nivå til ca. 20 %. Senere har resultatene bedret seg (Vox, 2015). Det er interessant å se om selve bruken av PC har innvirkning på resultatet. Den gamle og den nye prøven er ikke direkte sammenlignbare, nå kan man ikke stryke, men blir plassert på et nivå uansett, og oppgaver og tidsbruk er noe endret. Samtidig er det åpnet for at man kan melde seg opp til prøver selv på internett, ved hjelp av et egenvurderingsskjema og annet materiell (Vox, 2016e, s. 10). Elevene avgjør selv hvilket nivå de vil melde seg opp til, og kandidater som deltar i opplæring vil få veiledning av sin lærer, blant annet via eksempelprøver (se Vox, 2016a). Prøvene bygger på *Det felles*

europeiske rammeverket for språk (UDIR, 2011) som har en nivåinndeling av språkferdigheter, med A1 som det laveste, og C2 som det høyeste. Vox lager prøver for nivåene fra A1 til B2. Prøvene i skriftlig produksjon er laget slik at nivå A1 og A2 testes på samme prøve. Neste prøve tester nivå A2 og B1, mens den siste tester nivå B1 og B2. De fleste elevene som tar prøven i skriftlig produksjon havner på nivå A2 (ca. 55 %), mens omtrent like mange havner på nivåene A1 og B1, ca. 20 % på hvert nivå (Vox, 2016d). B2-prøve ble gjennomført for første gang i desember 2015. Elevene melder seg på én av prøvene, men oppgir ikke hvilket nivå de ønsker å nå, så det er ikke mulig å vite om de nådde målet sitt. I og med at elevene bare kan havne på de nivåene de har meldt seg på eller lavere (Vox, 2016e), er det nok en del elever som melder seg på «for høyt» nivå, kanskje av frykt for bli undervurdert.

Ved hjelp av teorier og tidligere forskning blant annet innenfor språktesting og andrespråksskriving har jeg brukt variabler for kompleksitet, nøyaktighet og flyt (CAF) for å sammenligne kvalitet i skriving i leksikon og syntaks i to skrivemoduser empirisk. Jeg har brukt kvantitative analysemetoder i dette arbeidet. Jeg har også vurdert om kravet til digital skriving i den nye norskprøven gjør at elever som ikke har tilstrekkelige ferdigheter på PC blir forfordelt, slik at konstruktvaliditeten ved den nye norskprøven dermed er svekket. I dette arbeidet har jeg særlig hatt et blikk på eldre voksne med antatt svakere ferdigheter i tekstbehandling. Endelig har jeg vurdert konstruktvalideringen av de nye norskprøvene.

1.1 Bakgrunn og avgrensning av tema

I Bibelen (12 Dom. 4:6) får vi høre om en språktest hvor tusenvis av efraimitter ble drept fordi de ikke uttalte ‘sjibbolett’ slik som gileadittene (efraimittene produserte sibilanten / s / initialt, der det skulle vært / ʃ /). Siden har ulike språklige forskjeller markert mer eller mindre uttalte skiller mellom inn-grupper og ut-grupper (McNamara, 2005). Begrunnelsen har i mange tilfeller vært politisk, og alltid verdiladet, sier Tim McNamara videre. Med nye verktøy og hjelpemidler oppstår gamle skikkelser ofte i ny drakt. I min jobb som lærer i voksenopplæring ble jeg interessert i å finne ut hvorfor andelen kandidater som oppnådde nivå B1 i skriftlig produksjon sank samtidig som Norskprøven ble digital. Det er klart at sammenhenger i menneskelige handlinger ofte er sammensatt. Det er sjelden én enkelt årsak til at noe skjer, og selv om to hendelser inntreffer samtidig eller like etter hverandre, er det ikke nødvendigvis et kausalforhold mellom dem.

Lærere i voksenopplæringen ser ofte at elevene har lite skrivetrening fra før. Særlig er argumenterende tekster og tekster som krever at eleven bruker fantasien noe nytt for mange elever. Videre vet vi at de tar med seg kulturkunnskap og kunnskap om skriving og sjangre fra tidligere lærte språk inn i andrespråket norsk (Selj, 2008). Det er også tydelig at ferdighetene i bruk av PC, og særlig bruk av PC som et skriveverktøy, er ujevnt fordelt blant elevene. Mange elever har hatt tilgang til PC i en del år, men da har de fleste brukt den til e-post, til internett, eller til å holde kontakt med venner og familie. De færreste har gjort bruk av PC til å skrive sammenhengende tekster. For noen deltakere er bruk av et skriveprogram i det hele tatt en ny opplevelse når de kommer på voksenopplæringen. Forskning gjort av blant andre Martha Pennington (2003) viser at bruk av PC kan gi fordeler både i skrivehastighet og antall feil, men at det kreves at eleven har fått opplæring i bruk av PC og tekstbehandling.

Deltakere som fikk rett og plikt til opplæring etter 1. september 2013 skal delta i norskopplæring i 550 timer. Dette er et minimum, og etter søknad kan de få inntil 3000 timer totalt. Personer som kommer fra EØS-området eller er over 67 år har ikke rett eller plikt til å delta i gratis norskundervisning. Det er min erfaring at de fleste deltakere på norskopplæring ikke er enspråklige når de kommer til Norge. Mange har allerede lært minst to språk, enten bare muntlig, bare skriftlig, eller både muntlig og skriftlig. Noen har kanskje ikke lært å skrive morsmålet sitt, mens andre ikke kan skrive noe særlig i det hele tatt, men kan snakke morsmål og ett eller flere andre språk i tillegg. Noen kan både snakke, lese og skrive godt på flere språk når de kommer til Norge. Graden av språklig bevissthet er også svært varierende, noen har ikke reflektert over språkets oppbygging, mens andre kjenner mange strukturer og regler på flere språk og kan overføre disse til språk de lærer senere. Av tidligere lærte andrespråk, er nok arabisk, engelsk og spansk mest framtredende. Slik ser vi at den språklige bakgrunnen til de forskjellige deltakerne er svært forskjellig.

Vox har utarbeidet *Læreplan i norsk og samfunnskunnskap for voksne innvandrere* (Vox, 2012). Her har de enkelte språkferdighetsnivåene (A1 til C2) *kompetansemål.* Disse er igjen oppdelt i *globale mål* og ulike *delmål*. Disse delmålene er *lytte, snakke, samtale, lese* og *skrive.* Språket som skal læres kjennetegnes ved *språklig bredde, uttale, flyt, grammatikk* og *kommunikasjon.* Arbeid med språket skal skje innenfor fire *domener* som skal gi innblikk i ulike sider ved det norske samfunnet: personlig domene, offentlig domene, opplæringsdomenet og arbeidslivsdomenet. I tillegg skal man altså jobbe med digital kompetanse integrert i de språklige kompetansemålene, som nevnt tidligere.

Organiseringen av voksenopplæringen er basert på *spor*, der spor 1 er tilrettelagt for deltakere som har liten eller ingen skolegang, og liten erfaring med skriftspråk. Dette sporet er igjen delt i en alfabetiseringsmodul og ordinær språkopplæring. Spor 2 er tenkt for deltakere som har en del skolegang og erfaring med skriftspråk, men ikke nødvendigvis med det latinske alfabetet. Spor 3 er tilrettelagt for deltakere som har god allmennutdanning, noen med erfaring fra høyskole- eller universitetsnivå. Disse har lang erfaring med skriftspråk og kanskje skolelærte fremmedspråk. Mange har gode læringsstrategier og høy språklig bevissthet (Vox, 2012).

En del kommuner har en relativt liten voksenopplæring, kanskje med én eller noen få klasser, sammensatt av deltakere fra flere spor. Deltakerne har en variert bakgrunn kulturelt og språklig, og ofte helt forskjellige mål og ønsker for opplæringen (Vox, 2011, s. 15). Noen ønsker å komme raskest mulig i jobb, noen skal lære å lese og skrive, mens andre skal forberede seg på videregående skole eller høyere studier. Alt i alt ser vi at tiden til å jobbe med skriving og digitale ferdigheter kan bli ganske begrenset for den enkelte deltaker. Det er naturlig å jobbe mot en avsluttende prøve, men det er også mye annet læreplanen setter som mål for deltakerne. Mange lærere føler at de har for lite tid til å jobbe med viktige emner, og i mange tilfeller kommer prøvene for tidlig i elevenes språklige utvikling. Elevene kan selv velge når de vil ta prøven, de kan melde seg opp via internett. Kandidater som kanskje tar prøven for tidlig, må ofte velge mellom å skrive raskt og risikere en del feil, eller å skrive sakte og få færre feil (Ellis & Barkhuizen, 2005, s. 143). Andre er opptatt av å variere språket sitt mest mulig eller skrive detaljert. Slike valg vil gi seg utslag i teksten, og ofte i vurderingen av prøven. I metodedelen vil jeg si mer om slike utslag i mitt prosjekt.

Norskundervisning og norskprøvene er slik at de gir deltakerne visse rettigheter og muligheter i samfunnet. Det kan være rett til å bli i Norge eller bli statsborger, muligheter til å få en jobb eller rett til å begynne på høyere studier. Stadig flere fylkeskommuner krever at voksne har B1-nivå for å komme inn i videregående utdanning (Pedersen, 2008, s. 99), og språkkravene øker i flere yrker (Dischler, 2011, s. 66). I et trangere arbeidsmarked kan man forvente at kravet til norskkunnskaper og dokumentasjon av dem blir høyere (Pedersen, 2008, s. 99). For mange elever kan prøven være en «high-stakes test», det vil si en viktig prøve som kan forandre folks liv og være vanskelig å gjøre på nytt (Bachman, 2004, s. 12).

Bak en prøve som gir viktige rettigheter til innbyggerne ligger det et stort arbeid med utvikling av prøven. Til det felles europeiske rammeverket for språk (CEFR) har *the Association of Language Testers in Europe (ALTE)* laget en *Manual for Language Test Development and Examining* (ALTE, 2011). Denne beskriver utviklingen av en språkprøve som en syklus hvor

man først bestemmer seg for at det er behov for en test, deretter utvikler testen, produserer testmateriale, gjennomfører testen, vurderer resultater, setter resultatene i en sammenheng og til slutt rapporterer resultatene til kandidater og andre som er impliserte (ALTE, 2011, s. 18). Etter hvert evalueres og revideres testen ved behov (ALTE, 2005, s. 46). En slik revisjon ble altså gjort for prøvene kom i ny utgave i mai 2014.

Norsk språktest er et samarbeid mellom Universitetet i Bergen, som har faglig ansvar, og Studieforbundet Folkeuniversitetet, som har det administrative og økonomiske ansvaret for å utvikle, gjennomføre og kvalitetssikre norskprøvene (Norsk Språktest, 2016). I en artikkel om utvikling av språkprøver, skrev Eli Moe ved Universitetet i Bergen for mer enn ti år siden at det er et krav om at språkprøver skal være greie og praktiske å gjennomføre, og at stadig flere (testorganisasjoner) derfor legger til rette for at prøver kan tas på datamaskin (Moe, 2003). Dette kan være én av grunnene til at bruk av PC kom inn med de nye norskprøvene. En annen grunn kan være såkalt *washback,* at undervisningen endres for å tilpasses prøven (Shohamy, 2001). Når skoler, lærere og elever vet at PC må brukes på prøven, vil det bli viktigere å bruke PC i undervisningen, noe som også reflekteres i læreplanen. Cecilie Carlsen (2007, s. 99) (som også er seniorrådgiver i Vox) antyder at det kan være hensiktsmessig å endre undervisningen ved å endre prøven.

Det er selvfølgelig også økonomiske motiver inne i bildet. I en rapport skrevet for Europarådet sies det direkte at antall tester forventes å øke, og at kostnaden per test må ned (Scheuermann & Björnsson, 2009, s. 7). Rapporten er ellers balansert og tar opp mulige problemer med validiteten. Og nettopp med tanke på validiteten er det interessant å se på overgangen fra papir til PC. Hvis de mentale prosessene som kreves for å svare riktig på en prøve endres ved at prøven tas på PC, kan validiteten til de inferensene vi gjør ut fra skåre på prøven også endres, sier Carol A. Chapelle og Dan Douglas (2006, s. 42). Med andre ord må validiteten til prøven vurderes på nytt.

En viktig del av utviklingsarbeidet med en ny prøve er som sagt å sikre at prøven måler det den sier den skal gjøre. Vi vil ikke at en prøve som sier den måler lesing også måler hvor godt deltakeren kan skrive. En *konstruktvalidering* er en kontroll av hvorvidt testen er en god operasjonalisering av det vi ønsker å måle (Pedersen, 2008). Dette bygger på et rammeverk utviklet av Samuel Messick (1989). Min oppgave kan være et lite bidrag til konstruktvalidering av de nye norskprøvene. Grant Henning (1991, s. 279) sier at god vurdering ikke er en statisk aktivitet, men er alltid mottakelig for tilpasning og forbedring. Det er gjort lite forskning på andrespråksskriving i det hele tatt i Norge, og særlig med et

kvantitativt blikk. Likevel går det an å hente innsikter både fra forskning i andre land og fra morsmålsforskning i Norge, og motsatt er det mulig å anvende metoder og innsikter fra min oppgave på forskning på morsmålselever.

1.2 Videre avgrensning, forskningsspørsmål og hypoteser

For å finne forskjeller mellom de to måtene å besvare prøven på, med penn og papir og med PC, var jeg interessert i å se på et antall prøver skrevet på hver måte og sammenligne dem. Jeg ønsket i utgangspunktet å se på autentiske prøver, men dette fikk jeg avslag på hos Norsk Språktest[1]. Begrunnelsen var at prøvene ikke kunne sammenlignes. Jeg valgte da å bruke et alternativ, å arrangere to prøver som lignet mest mulig på norskprøvene. Ved skolen jeg arbeidet var det noen klasser som hadde elever på nivå A2/B1, så jeg satte sammen to balanserte grupper som tok prøven i hver sin modus. Ved hjelp av et utvalg leksikalske og syntaktiske variabler og noen statistiske tester undersøkte jeg om kvaliteten på skrivingen ble endret ved overgangen til PC. Elevenes forskjellige bakgrunn, og kanskje særlig deres alder og erfaring med bruk av PC som et skriveverktøy, kunne ha en betydning for om de skrev best på PC eller med penn og papir. Jeg har hentet teori og innsikter fra forskning i andre land og Norge, på felter som andrespråksskriving, skriving på PC, testing og konstruktvalidering.

Jeg så på skriveredskapets påvirkning på skriftlige andrespråksteksters kvalitet, med et særlig blikk på eldre kandidater med svake ferdigheter i tekstbehandling. Jeg så også om konstruktvalideringen i de nye norskprøvene har vært god nok.

Jeg arbeidet ut fra følgende forskningsspørsmål og hypoteser:

- Med bakgrunn i en hypotese om at kvaliteten i skriftlige andrespråkstekster øker når deltakerne skriver på PC, stilte jeg forskningsspørsmålet: Hvordan og i hvilken grad forandres kvaliteten i tekstene, målt i syntaks og ordforråd, når andrespråksskrivere skriver på PC?
- En hypotese om at alder og erfaring med bruk av PC som skriveverktøy virker inn på tekstenes kvalitet, gir forskningsspørsmålet: I hvilken grad spiller alder og antatt erfaring med bruk av PC som skriveverktøy en rolle i kvaliteten på skriftlige tekster?

[1] Personlig kommunikasjon, e-post fra Cecilie Hamnes Carlsen ved Norsk Språktest 6.7.2015. Se også Vox (2016f).

- Forskning forteller at elever med tilstrekkelige PC-ferdigheter skriver bedre tekster på prøver med tidsbegrensning fordi de har mer tid til planlegging og revisjon når selve transkriberingen tar kortere tid. I hvilken grad gjelder dette for andrespråkselever?
- I denne sammenhengen er det naturlig å anta at de nye norskprøvenes konstruktvalidering ikke har vært god nok. Et spørsmål til dette blir da: I hvilken grad er det datakunnskaper og ikke språkkunnskaper som vektlegges i de nye norskprøvene?

1.2 Oppsummering og videre framstilling

Fra 2013 ble Norskprøven obligatorisk for innvandrere til Norge som har rett og plikt til norskundervisning. Prøven i skriftlig framstilling har blitt digitalisert, men mange elever har svake ferdigheter i bruk av tekstbehandling. Nivåene i Norskprøven kan gi elevene rettigheter og muligheter, så det er viktig at den valideres når den endres. Fordi jeg ikke fikk tilgang til autentiske tekster i undersøkelsen min måtte jeg finne en alternativ framgangsmåte. For å finne forskjeller i kvaliteten i skriftlig produksjon hos voksne andrespråksskrivere gjorde jeg en kvasieksperimentell undersøkelse med to balanserte grupper av deltakere, hvor den ene gruppa skrev med penn og papir og den andre gruppa skrev på PC. Jeg anvendte noen syntaktiske og leksikalske variabler på tekstene og sammenlignet disse statistisk. Jeg undersøkte om tekstene ble bedre når elevene skrev på PC. Videre undersøkte jeg om alder og antatt tidligere erfaringer med PC som skriveverktøy hadde en innvirkning på resultatene. Jeg så også om elever med tilstrekkelige ferdigheter på PC skrev bedre tekster fordi de fikk mer tid til planlegging og revisjon. Undersøkelsen av de tre første spørsmålene kan være ledd i en konstruktvalidering av de nye norskprøvene.

I kapittel to vil jeg gjennomgå teorier og tidligere forskning på andrespråksskriving, skriving på PC og språktesting. Kapittel tre er et metodisk kapittel som dreier seg om hvordan jeg konstruerte og gjennomførte testen og om deltakerne i den. I kapittel tre vil jeg også se på de avhengige variablene i avhandlingen, det vil si måleenheter for kvalitet i skriving, og hvordan disse er utviklet og anvendt tidligere. Også statistiske metoder og valg i den forbindelse vil bli drøftet i kapittel tre. I kapittel fire viser jeg resultatene fra undersøkelsen. Disse blir drøftet i kapittel fem, før jeg antyder noen implikasjoner til slutt.

2. Teoretiske perspektiver og tidligere forskning

Denne oppgaven plasserer seg innenfor forskningsfeltet *norsk som andrespråk.* I en artikkel fra 2007 deler Anne Golden, Lise Iversen Kulbrandstad og Kari Tenfjord forskningen innenfor dette feltet i tre utviklingslinjer, innlærerspråk, språk- og kulturkontakt og didaktikk. De plasserer arbeider knyttet til evaluering og testing i den didaktiske linjen (Golden, Kulbrandstad & Tenfjord, 2007). Min oppgave er hovedsakelig didaktisk, men berører også innlærerspråk. Måling er både data- og teoridrevet, sier Messick (1987, s. 43f), og er egentlig den empiriske koblingen mellom fenomener som kan forskes på og de teoretiske påstandene forskere vil framsette om disse fenomenene (Norris & Ortega, 2009, s. 557).

I dette kapittelet vil jeg redegjøre for noen av disse teoriene og tidligere forskning. Først vil jeg definere begrepene andrespråk og mellomspråk, før jeg ser på skriveprosessen på morsmål og i et andrespråk. Så vil jeg vise hvordan andrespråksskriving oppsto som et fagområde og hvordan det kategoriseres og defineres før jeg sier noe om andrespråksskriving i Norge. Deretter vil jeg drøfte hvordan skriving på PC skiller seg fra skriving med penn og papir, med et særlig fokus på eldre voksne og innvandrere. Utviklingen av språktesting vil bli gjenstand for drøfting, der jeg vil ha et særlig fokus på konstruktvaliditet i en moderne forstand av begrepet. Jeg vil også vise hvordan jeg posisjonerer meg i denne oppgaven innenfor fagets rammer. Alt dette vil skje med bakgrunn i tidligere forskning og teorier som er utviklet innenfor fagområdene.

2.1 Andrespråk og mellomspråk

Deltakerne i voksenopplæring er grovt sett mellom 16 og 55 år[2], og har begynt å lære seg norsk i voksen alder. De yngste har kanskje vært i Norge et år eller to, og begynte på voksenopplæring da de var 16 år gamle. Innvandrere under 16 år går vanligvis på ordinær ungdomsskole, eventuelt tilpasset med særskilt norskopplæring. Alle har da norsk som et andrespråk. Begrepet *andrespråk* skal ikke tas så bokstavelig, det betyr at det er tilegnet etter førstespråket, altså morsmålet. Det er da et rekkefølgefenomen, og kan altså være språk nummer to, tre, fire og så videre (Abrahamsson, 2009, s. 13; Berggreen & Tenfjord, 2007, s.

[2] Deltakere mellom 55 år og 67 år har rett, men ikke plikt til deltakelse. I 2014 var det 415 deltakere over 55 år i voksenopplæring i Norge (SSB, 2016).

16). I praksis vet vi at svært mange deltakere har både ett og to andrespråk før de kommer til Norge, men alle språk som læres etter førstespråket kalles andrespråk.

I forskningen på andrespråk er tanken om at innlærere utvikler et eget språk mellom morsmålet og målspråket, et mellomspråk, ukontroversiell. Selv om flere forskere beskrev dette parallelt, ble det Larry Selinkers betegnelse *interlanguage* som ble stående, også fordi han prøvde å utvide det kontrastive perspektivet og beskrive mellomspråket som en kognitiv prosess (Hammarberg, 2013, s. 30). Mellomspråk, særlig i en tidlig fase, er enkle, ustabile og variable språk som viser tverrspråklig innflytelse, vi finner «…spor etter elementer, enheter, regler eller mønster fra innlærerens morsmål» sier Harald Berggreen og Kari Tenfjord (2007, s. 29). Dette mellomspråket har sitt eget system, en grammatikk, hos innlæreren, og er ikke en feilvariant av målspråket. Innlæreren prøver ofte ut hypoteser om språket, vi kan derfor ofte finne flere varianter av samme målspråksform i mellomspråket (Berggreen & Tenfjord, 2007, s. 301). Men målspråket er ikke alltid lett å få tak i. Et problem med vurdering av mellomspråk er at målspråket ikke er en skriftlig standard med allmenn konsensus, sier Barbara Kroll (1990, s. 141) om engelsk. Norsk har sannsynligvis enda større variasjon i skriftlig valgfrihet og er dermed mindre transparent enn engelsk.

2.2 Skriveprosessen

2.2.1 Førstespråk

Empirisk forskning på skriveprosessen, i motsetning til produktet, begynte med Janet Emigs *The Composing Processes of Twelfth Graders*, der hun beskrev prosessen som tredelt, bestående av planlegging, selve skrivingen (transkribering) og revisjon. Denne prosessen foregår ikke lineært, men rekursivt, altså at deler av prosessen, eller subrutiner, gjentas hele tiden (Nystrand, Greene, & Wiemelt, 1993). Påvirket av dette utviklet John R. Flower og Linda Hayes i 1981 en kognitiv modell av skriveprosessen, som består av de tre nevnte komponentene (med underkomponenter), disse styres av en *monitor*. De tre komponentene samarbeider med langtidsminnet, og det hele tar utgangspunkt i skrivesituasjonen (Flower & Hayes, 1981). Planlegging skal ikke forstås som planlegging forut for oppgaven, men som å få en idé eller et «bilde» i hodet. Skrivingen blir da en «oversettelse» av idéen eller bildet til skrift (Fjørtoft, 2014). Modellen er senere utvidet av Hayes med sosiale og motivasjonelle faktorer.

2.2.2 Andrespråk

Blant teorier om skriveprosesser på andrespråket vil jeg nevne én som et eksempel. Ernesto Macaro (2003) er påvirket av modellen til Flower og Hayes. Han foreslår en prosess delt i to interne komponenter (arbeidsminne og langtidsminne) og tre ytre komponenter (oppgavekrav, ressurser og det skriftlige produktet i øyeblikket). Mellom disse har han seks rekursive prosesser, der den ene involverer andrespråket på flere måter (Macaro, 2003, s. 222f). Teorier om skriveprosessen i andrespråket har hatt liten betydning i forskning, sier Charlene Polio (2012), og hevder at dette skyldes at oppgaven med å skape en slik modell er for stor. Modellen er komplisert nok i førstespråket, sier hun, om en ikke skal ha et andrespråk inn i tillegg. Polio nevner flere teorier som har vært foreslått, men de mangler et tydelig bilde av den lingvistiske kunnskapskomponenten og en kobling til andrespråksteorier (Polio, 2012). Det er ikke vanskelig å være enig med Polio i at dette er komplisert, som i Macaros modell over. Når eleven skal skrive på et andrespråk som han ikke behersker, og møter ukjente kulturelle og retoriske krav og nye oppgavetyper, er det klart at modellen må ta inn andre komponenter enn modeller for førstespråk. Vi vet at teorier om andrespråk sier at mellomspråket er under utvikling, og at elevene i ulik grad er avhengig av førstespråket. Dermed er det også mulig det ikke er nok med én modell. Poenget i min sammenheng er at den kognitive belastningen er høyere når man skriver på et andrespråk enn på førstespråket, iallfall relativt tidlig i andrespråkstilegnelsen (Selj, 2008). Lourdes Ortega (2009, s. 236) hevder at omkring 60 % av skrivetiden går med til formulering (transkribering) når prøven er begrenset på tid, resten av tiden fordeles på planlegging og revisjon. Eleven må over et terskelnivå i andrespråket for å få mer balanse mellom de tre prosessene, sier hun. Sara Cushing Weigle (2002, s. 35) støtter dette siste. Det virker sannsynlig at andrespråkselever bruker mindre tid på planlegging og revisjon enn førstespråkselever på en prøve.

I det foregående har jeg vist hvordan andrespråk og mellomspråk defineres og at skriveprosessen er en komplisert kognitiv prosess. Jeg har også vist at teorier om skriveprosessen på et andrespråk finnes, men at de ikke har hatt stor betydning i forskningen. Likevel er det sannsynlig at det er mer krevende å skrive på et andrespråk enn på morsmål, og at tiden til planlegging og revisjon blir kortere. I det neste avsnittet skal jeg se på forskningen på andrespråksskriving og hvordan andrespråksskriving skiller seg fra skriving på morsmål. Jeg vil også vise hvordan andrespråksskriving kan kategoriseres, og hvordan jeg posisjonerer meg innenfor disse kategoriene.

2.3 Forskning på andrespråksskriving

Forskning om skriving på et andrespråk er et relativt nytt område. Skriveforskeren Paul Kei Matsuda (2003) gir fokuset på den audiolingvale retningen og talespråkets framtredende stilling innenfor anvendt lingvistikk ansvaret for at forskning på andrespråksskriving ikke startet før på 1960-tallet. Da begynte store grupper andrespråksstudenter på nordamerikanske universiteter, og man så behovet for mer kunnskap. En viktig periode i fagets historie var tidlig på 1990-tallet, da en rekke bøker kom ut, det ble avholdt egne konferanser innenfor feltet, og særlig at Journal of Second Language Writing ble etablert i 1992 med en rekke artikler innenfor underkategorier av andrespråksskriving. Fagfeltet oppsto i skjæringspunktet mellom andrespråksstudier og skrivestudier (composition studies), sier Matsuda (2003), og forskere fra begge «leire» har litt forskjellige innganger til feltet.

I et kapittel om skriving på andrespråk beskriver Ulrika Magnusson andrespråksskriving som et særskilt aspekt av andrespråksutvikling, delvis avhengig av det allmenne andrespråksnivået, men også av andre faktorer, blant annet alder og lese- og skrivekompetanse på førstespråket (Magnusson, 2013). Det er umulig å unngå en viss grad av spekulasjon når det gjelder koblingen mellom andrespråkteorier og skriving, sier Polio (2012). Alder ved ankomst til målspråkslandet er ett av områdene det er diskusjon om, innlærere som ankommer tidlig når et høyere nivå enn de som ankommer senere (Abrahamsson & Hyltenstam, 2013), men eldre innlærere kan dra fordeler av kunnskap om tekst, lesing og skriving på førstespråket som kan overføres til andrespråket (Magnusson, 2013). Innlærere har i mindre grad enn innfødte målspråksbrukere kunnskap om sjanger og stil (Kaplan, 1966; Magnusson, 2013; Selj, 2008). Fagfeltet kalt *kontrastiv retorikk* gir grovt sett tre forklaringer på dette: lingvistiske, kulturelle og utdanningsmessige (Matsuda, 1997). Særlig er utviklingen av skolerelatert eller akademisk språkbruk tidkrevende (Cummins, 1981, 2008; Magnusson, 2013).

I en sammenligning av 72 studier av skriving på første- og andrespråk fant Tony Silva (1993) at andrespråksskriving skilte seg fra førstespråksskriving på flere måter. I de fleste av studiene var elevenes skrivetid begrenset til 30 til 60 minutter. Generelt planla andrespråksskrivere mindre, sier Silva, og de satte seg færre mål, både lokalt og globalt i teksten. De nådde heller ikke målene de satte, og hadde vanskeligere for å organisere generert materiale. Selve skrivingen (transkriberingen) var mer strevsom, mindre flytende og mindre produktiv, hevder Silva. Andrespråkselevene brukte mer tid på å se på oppgaveteksten og i ordbok, og var mer

opptatt av ordforrådet. De hadde også flere og lengre pauser. Elevene reviderte mindre og reflekterte mindre over skrevet tekst, men en del av revisjonen ellers syntes lik som i førstespråket, selv om noen av funnene på revisjon sprikte. Revisjonen fokuserte på grammatikk, og mindre på staving (Silva, 1993, s. 661f). De tekstlige trekkene viste at elevene gjorde flere feil totalt, både i syntaks og leksikon, og ble rangert som svakere i holistiske vurderinger (Silva, 1993, s. 663).

Innlærere pakker informasjonen mindre, bruker et mer ensartet vokabular og utvikler diskursen mindre, sier Dudley W. Reynolds (2005), dessuten involverer de ikke leseren så godt i teksten, og forklarer sine utsagn dårligere, sier han. Innlærere har mindre bredde i ordforrådet, det vil si at de behersker færre ord, både reseptivt og produktivt, og har mindre dybde, de kjenner gjerne bare én betydning av ord som er homonymer eller homografer, viser Batia Laufer (1998) og Åke Viberg (2004). Videre finner Siok H. Lee (2003) at innlærere oftere bruker uformelle, høyfrekvente ord, og få lavfrekvente. Disse funnene på leksikonnivå støttes av Scott A. Crossley og Danielle S. McNamara (2009) og Judit Kormos (2011). Funnene på syntaktisk nivå støttes av Puangpen Intaraprawat og Margaret S. Steffensen (1995), Eli Hinkel (2003) og Judit Kormos (2011).

Andrespråksinnlærere overfører kunnskap om språk fra tidligere lærte språk til andrespråket, som nevnt. Alexander Friedlander (1990) finner i et doktorgradsprosjekt at også kunnskap om *skriving* (eller mangel på den) overføres fra førstespråket. Dette imøtegås av flere, Joan E. Carson og Phyllis E. Kuehn (1992) viser at overføringen ikke nødvendigvis gjelder svake morsmålsskrivere, og Mohammad Aliakbari (2002) finner at korrelasjonen mellom skriving i førstespråk og andrespråk er så lav at andre faktorer er involvert. Til støtte for Friedlanders syn hevder Alexandra Rowe Krapels (1990) i en gjennomgang av tidligere forskning at kunnskap om skriving er viktigere enn lingvistisk kunnskap. Anne Lene Berge (1999) tar opp norsk som andrespråk i videregående skole, og viser at mange andrespråkselever har problemer med å tilpasse seg våre tekstkonvensjoner, det kan være linearitet, objektivitet, mottakertilpasning, tekstlig makrostruktur og krav til kritisk refleksjon rundt innholdet (A. L. Berge, 1999, s. 141). Mange voksne andrespråksskrivere oppnår aldri målspråksferdighet, sier Ken Hyland, (2003, s. 32), dette skyldes enten at de når et kompetansenivå som de kan kommunisere tilfredsstillende på, eller at språket «fossiliseres» på et visst nivå. Mye av dette skyldes individuelle forskjeller, fortsetter han (om fossilisering se Long, 2003, s. 487). Hyland (2003) advarer mot å se innlærerne som en homogen gruppe, alle har ulikt startsted og

progresjon fram til utviklingen flater ut. Dette støttes av Sofie Johansson Kokkinakis og Ulrika Magnusson (2011), som ser forskjellen mellom første- og andrespråksbrukere mer som et kontinuum enn en binær motsetning.

Polio (2003) viser i en oversiktsartikkel at forskningen innenfor andrespråksskriving kan deles i fire: Studier som fokuserer på skrivernes *tekst*, det vil si *produktet*, studier som fokuserer på *skriveprosessen*, studier som fokuserer på *deltakerne* i prosessen, og endelig studier som fokuserer på *konteksten* for skrivingen, både i og utenfor klasserommet (Polio, 2003). I denne sammenhengen vil min oppgave berøre alle disse fire, men i all hovedsak vil tekstene være det viktigste, med deltakerne som nummer to. Fordi det meste av forskningen har som mål å hjelpe skriverne til å skrive bedre tekster, er det ikke rart at det er forskningen på tekst som er størst, sier Polio (2003). Forskere som vil kvantifisere kvaliteten i skrivingen har gjort noen valg, hevder hun videre. Noen har valgt et holistisk mål som gir en skåre til teksten som helhet, andre har valgt en analytisk skala som er sammensatt av delskårer på forskjellige aspekter ved teksten. Innhold, koherens og diskurs kan vurderes i en tredje variant. Den siste muligheten er å velge mål for kompleksitet, nøyaktighet og flyt. Det er det jeg har gjort, og jeg vil si mer om det i metodekapittelet.

Andrespråksskriving kan defineres på ulike måter, sier Anne Golden og Rita Hvistendahl (2015). Den første er ut fra *skriverens språkbakgrunn.* En slik definisjon vil gi andrespråkstekster selv om teksten er målspråkslik, og er temmelig statisk. En annen og mer dynamisk definisjon er å ta utgangspunkt i *tekstene*, altså at tekstene viser *typiske innlærertrekk.* Særlig avvik på de laveste språknivåene tillegges vekt, og ortografiske, morfologiske, syntaktiske og leksikalske avvik regnes blant disse, sier Golden og Hvistendahl (2015). Disse avvikene er ikke forbeholdt lavere trinn i skolen, men kan henge med opp i høyere utdanning og i arbeidslivet.

Andrespråk som forskningsobjekt i Norge oppsto rundt 1980, forteller de videre, og siden en del av dataene i hovedoppgavene som kom på denne tiden var fra innlærertekster, markerer disse hovedoppgavene også starten på forskning på andrespråksskriving i Norge. Forskningen på dette i Norge deler Golden og Hvistendahl i fire kategorier: studier på språknivå, studier på tekstnivå, studier av språkopplæringen og skriving som sosial praksis. Antall arbeider varierer en del, men til 2015 var det under ti hoved- og masteroppgaver som har sett på selve tekstene, så det internasjonale fokuset på tekstene gjenspeiles ikke her hjemme. Av norske studier på tekstnivå, har de fleste sett på tekstbinding eller uttrykk for logiske relasjoner, viser Golden

og Hvistendahl (2015). For snart ti år siden sa Olga Dysthe og Frøydis Hertzberg at «Det flerkulturelle perspektivet må bli en selvfølgelig del av skriveforskningen, og det samme må digital skriving…» (Dysthe & Hertzberg, 2007, s. 22). Også Kjell Lars Berge og Johan L. Tønnesson påpeker at digital og minoritetsspråklig skriving bør være satsingsområder framover (K. L. Berge & Tønnesson, 2007). Det har ikke vært noen stor økning av publisert forskning på andrespråksskriving likevel. Ikke før i februar 2016 disputerte Ingrid Jølbo (2016) over den første doktorgradsavhandlingen i andrespråksskriving i Norge. Sverige var tidligere ute, og har hatt større produksjon, Danmark ligger noe foran Norge. Med min avhandling skriver jeg meg rett inn i det som Dysthe og Hertzberg og Berge og Tønnesson ønsker seg av skriveforskning. Jeg har ikke funnet noen norske arbeider som har sett på kvaliteten i andrespråksskriveres tekster på flere språklige nivåer og gjort et forsøk på å kvantifisere den, slik jeg gjør i denne oppgaven.

I dette avsnittet har jeg vist framtredende trekk ved andrespråksskriveren og historikk, kategorisering og definisjon av andrespråksskriving som et eget fagområde. Jeg har også vist hvordan jeg har posisjonert meg innenfor kategoriene i andrespråksskriving i denne oppgaven. Jeg vil fokusere på skrivernes tekst, altså det skriftlige produktet. Jeg vil også velge noen variabler for kompleksitet, nøyaktighet og flyt for å se på kvaliteten i disse tekstene. Dette vil jeg si mer om i metodekapittelet. I det neste avsnittet vil jeg definere skriving på PC, og hvordan slik skriving skiller seg fra skriving med penn på papir. Jeg tar utgangspunkt i generelle studier på elever, men fokuserer etter hvert mer på eldre voksne og innvandrere, som har særlige utfordringer med skriving på PC.

2.4 Å skrive på PC

Skriving på PC eller tekstbehandling er et eget fagfelt med start rundt 1985 (van Waes, 1994). Studier som sammenligner skriving på PC og skriving med penn og papir var vanligst på slutten av 1980-tallet, men antallet har sunket mye siden (MacArthur, 2006, s. 250). Forskningen har blitt mer vridd mot skriving *på internett* (CMC) og *hvordan* skriving på PC best kan gjøres. Det kan også ha sammenheng med at den digitale teknologien har gjort sitt inntog i nær sagt alle domener, og at elevene har vokst opp med å skrive på PC eller andre digitale verktøy. Også lærerne har akseptert PC, om enn i varierende grad, og antall maskiner i skolen har økt betydelig. Margareth Sandvik (2008) sier at digital kompetanse kan operasjonaliseres i en rekke delferdigheter på ulike nivåer. Disse er både grunnleggende

ferdigheter og mer avanserte og kritiske ferdigheter, blant annet å kunne skape tekster ved hjelp av digitale verktøy. Digital kompetanse er et samspill mellom tekniske, kritiske og tekstuelle ferdigheter, sier hun (Sandvik, 2008, s. 159f). En PC har noen fordeler framfor andre digitale verktøy når det gjelder tekstbehandling. Fordi den har muligheter for tekstskaping, revisjon og flytting, mener Charles A. MacArthur (2006) at den passer godt til den rekursive skriveprosessen. Sandvik hevder videre at vi står foran en ny generasjon av elever som er «multikulturelle, «multilingual» eller flerspråklige, og de er multimediale og multimodale – de har med andre ord høy grad av digital kompetanse» (Sandvik, 2008, 167f). Hun har sett på elever i barneskolen. Og bruken av PC i samfunnet har økt mye de siste årene. Andelen av personer i Norge som har brukt PC daglig har økt fra 45 % i 1995 til 95 % i 2014, og i skolen har det skjedd en tredobling av PC-bruken, fra 8 % til 24 % i samme periode (Medienorge, 2016).

Studier som sammenligner skriving på PC og skriving med penn og papir finner varierende resultater. En tidlig metastudie (N = 8) rapporterte entydig positive resultater både for lengde og kvalitet ved bruk av PC, men ikke for revisjon (Roblyer, Castine, & King, 1988). Robert L. Bangert-Drowns (1993) viser to litt senere metastudier (Cochran-Smith et al., 1991 og Russell, 1991) som fant svært små effekter av PC. Bangert-Drowns (1993) så på 33 studier hvor to grupper av elever fikk lik instruksjon, men skrev i forskjellig modus. I 17 av studiene var elevene i college, resten var yngre elever. Holistisk vurdert økte kvaliteten signifikant med PC, særlig for svakere elever (forklart med økt motivasjon), og lengden i antall ord økte. Bangert-Drowns fant også at kvaliteten økte selv om lengden økte. Han fant ingen entydige resultater for revisjon. Bangert-Drowns (1993) hevder at kognitiv kapasitet frigjøres fra enklere oppgaver (mekaniske) ved bruk av PC slik at mer kapasitet kan gå til tekstskaping. Dette er selvfølgelig avhengig av at tastaturferdighetene er automatiserte og effektive. Bangert-Drowns sier at mange faktorer innvirker på skriving, blant annet kontekst, men særlig er det viktig med instruksjon på det som gir PC en fordel framfor penn og papir, sier han.

Luuk van Waes (1994) viste i en metastudie av forskning gjort fram til 1993 med 40 studier i hver modus at også rettskriving og tegnsetting ble forbedret med PC. I en metaanalyse av 26 studier mellom 1992 og 2002 gjort på elever i de første tolv årene av utdanningen (K-12) fant Amie Goldberg, Michael Russell og Abigail Cook (2003) signifikante forskjeller mellom skriving på papir og på PC i favør av PC, både i kvantitet og kvalitet. De hevdet at dette delvis skyldtes samarbeid og repetisjon. Studier som fokuserte på revisjon viste ingen tydelig

forskjell. MacArthur (2006) hevder at revisjon måles på forskjellige måter, både underveis og mellom utkast, og at dette er årsaken til at studiene ikke finner en bedring. En PC legger til rette for revisjon, men ofte bare på den delen av teksten som vises på skjermen, slik at revisjonen blir på lokalt nivå. Elever trenger opplæring i global revisjon, sier Pennington (2003) og Jocelyne Bisaillon (1999), og bruken av PC alene er ikke nok.

I en ganske ny studie sammenlignet Penelope Collins, Jin Kyoung Hwang, Binbin Zheng og Mark Warschauer (2013) to grupper elever, den ene gruppa skrev på PC, den andre med penn og papir. Elevene var i alderen 10 til 12 år og hadde forskjellige morsmål. Gruppa som skrev på PC skapte signifikant bedre tekster, målt i ord, tekstlengde og kompleksitet, og holistisk vurdert. Dette skyldes nok delvis at elevene fikk undervisning i å skrive på PC og hadde sin egen PC i alle skoletimer. Videre fikk de tilgang til stavekontroll og ordbok, som kontrollgruppa ikke fikk. Lærerne og sensorene visste også hvem som skrev i hvilken modus.

På spørsmålet om bruk av tekstbehandling virker, svarer Mac Arthur i en oppsummering av forskningen: «It depends». Alene har tekstbehandling minimal innvirkning, sier han, men kombinert med instruksjon kan elevene utvikle bedre skriveferdigheter (MacArthur, 2006, s. 260). Jeg tror også at det er vanskelig å lage metaanalyser i dette feltet, da det eneste disse studiene har felles, er skriveredskapet (se Clark (1985) for en diskusjon om dette). Kontekst, elevenes alder, samarbeid, tilbakemelding, tid, oppgavetyper, hjelpemidler, måleenheter og andre faktorer varierer, da er det krevende å si noe generelt.

Ylva Hård af Segerstad og Sylvana Sofkova Hashemi (2006) hevder at talelignende grammatiske strukturer og ortografi, manglende tegnsetting og bruk av stor bokstav kan være påvirket av uformell bruk av digital skriving i fritida. De mener også at frekvent bruk av forkortelser kan gjøre at de blir konvensjonalisert i alle typer digital tekst. Naomi Baron (2008) hevder at dette er mindre vanlig enn vi tror, og heller ikke noe nytt. Slike strukturer i mine besvarelser vil føre til et økt antall feil i den grad de ikke er konvensjonalisert i skrift. Jeg drøfter feil mer i metodekapittelet. Å skrive på PC innebærer naturligvis også å lese på PC-skjerm. På dette er det gjort en rekke studier, også i Norge. Det vil føre for langt å gå inn på denne problemstillingen her.

2.4.1 Når eldre og innvandrere skriver på PC

Vi så at mange i Norge bruker PC mye mer enn før. Det er usikkert hvor mange av kandidatene til Norskprøven dette gjelder. Noen har god erfaring med bruk av digitale verktøy, mens andre knapt har brukt slike verktøy. Alder, utdanning, opprinnelsesland og sosioøkonomiske faktorer spiller inn her (McNamara, 2000, s. 80f og 118), og selvfølgelig hvor mye tid kandidaten har brukt på dette hjemme og på skolen før prøven. I USA kan skillet mellom fattige innvandrere uten leseferdigheter og hvite morsmålsbrukere være like stort som skillet mellom USA og utviklingsland («Digital Divide») når det gjelder PC-bruk, sier Mark Warschauer og Meei-Ling Liaw (2010). Dette skillet er nok ikke like stort i Norge, blant annet fordi Norge er et mer egalitært samfunn. Også Edward W. Wolfe og Jonathan R. Manalo (2004) finner at minoriteter, og særlig eldre, har mindre tilgang til og erfaring med PC enn majoriteten (i USA). Hvis morsmålet har ikke-vestlig alfabet kan det også spille inn når de skal skrive på en PC med vestlig alfabet. Eldre og innvandrere har også mer frykt for å bruke PC, og det er mindre sjanse for at de vil velge å bruke PC hvis de får et valg, hevder de videre.

Ferske tall for PC-bruk i verdens land er vanskelig å finne, men vi kan anta at bruk av internett er representativt. Da finner vi at mens 98 % av befolkningen i Norge bruker internett (juli 2016) er det 77 % i Litauen, 72 % i Polen, 71 % i Russland, 59 % i Bulgaria, 44 % i Ukraina, 43 % på Filippinene og Thailand, 30 % i Syria, 20 % i Indonesia, 13 % i Irak, 7 % i Afghanistan, 2 % i Somalia og 1 % i Eritrea som gjør det, ifølge nettsiden internetlivestats.com (2016)[3]. Selv om vi føler oss omgitt av internett, er det fortsatt 54 % av verdens befolkning som ikke har slike muligheter. Disse tallene gjelder tilgang til internett på alle plattformer, det vil si mobiltelefon, nettbrett, PC med mere (ibid.). Skriveferdigheter på flere av disse plattformene kan derfor ikke overføres til PC uten videre. Det går selvfølgelig an å skrive på datamaskin uten å ha internett, poenget her er den store forskjellen mellom Norge og mange andre land når det gjelder tilgangen til internett, og dermed sannsynligvis til datamaskiner. I en rapport for ETS[4] der nesten 134 000 tekster fra 200 land ble undersøkt viser Wolfe og Manalo (2005) at sosial bakgrunn hos elever slår ut på prøver tatt med penn og papir. Det er interessant at de hevder at denne bakgrunnen slår ut mer ved testing på PC ved at også

[3] Tallene er avrundet til nærmeste hele prosent.

[4] Educational Testing Service er en av de tre største private testorganisasjonene i USA (Spolsky, 2014, s. 1577).

mulighetene for *å lære seg* bruk av PC er mindre. Jan A.G.M. van Dijk (2006) hevder også at «gamle» ulikheter mellom sosiale grupper når det gjelder ressurser og kapital forsterkes med digitale media.

Pennington har gjort en rekke studier på voksne andrespråksskriveres skriving på PC, og hevder at skriveprosessen forenkles ved bruk av PC, slik at mer fokus kan gå til tekstproduksjon. Hun advarer likevel mot at eldre, og mennesker med «computer-phobia» ikke har de samme fordelene. Hun sier at positive holdninger til PC påvirker skrivingen, noe som gir bedre kvalitet og større kvantitet, men at negative holdninger eller usikkerhet kan gi den motsatte effekten (Pennington, 2003). Ellers finner hun de samme generelle effektene som nevnt i metastudiene for yngre elever. I 1991 var bruk av PC ganske nytt for de fleste, og kan kanskje sammenlignes med hvordan eldre og uerfarne opplever PC-bruk i dag. Joyce Neu og Robin Scarcella (1991) så på holdninger til PC hos 54 andrespråksstudenter (alderen er ikke oppgitt) som skulle skrive på engelsk. De aller fleste lærte PC-bruk fort, de syntes det var nyttig i flere deler av skriveprosessen og de fikk bedre resultater. Likevel er det interessant å legge merke til at 10 % sa at de var nervøse når de skrev på PC, 22 % hadde vanskeligheter med å forstå hvordan maskinen skulle brukes, og 20 % var redde for å ødelegge datamaskinen (Neu & Scarcella, 1991). van Dijk (2006) hevder at slike problemer ikke forsvinner selv med økt bruk. Ruru S. Rusmin (1999) har sett på eldre andrespråksskrivere, og sier at det tar tid å tilpasse skrivestrategier eller skape nye samtidig som en skal lære teknologien. Rusmin sier også at noen kan mislike datamaskiner, og at dette påvirker holdninger, prosess og resultat når det gjelder skriving på PC. Eldre voksne bruker lengre tid på å lære tekstbehandling og gjør flere feil når de jobber med tekstredigering (Dyck & Al-Awar Smither, 1996, s. 108).

Disse funnene ble gjort for en del år siden, men senere fant Nancy Horkay, Randy Elliot Bennett, Nancy Allen, Bruce Kaplan og Fred Yan (2006) at elever i Grade 8 (alder 13-14 år) skrev signifikant dårligere på PC hvis de ikke hadde tilstrekkelig kompetanse på tekstbehandling. Over 90 % av disse elevene oppga at de brukte PC hjemme, og 87 % sa at de brukte PC en del på skolen. Elevene brukte likevel ikke *tekstbehandling* nok til å få en tilstrekkelig skrivehastighet. Horkay et al. (2006) sier at disse heller ikke nødvendigvis er dårligere til å skrive, men de er dårligere til å skrive *på PC*. De hevder videre at man undervurderer elevenes skriveferdigheter hvis man bare gjennomfører skrivetesten i én modus. Wolfe og Manalo (2005, s. 11) hevder at skriving på PC for kandidater med lave ferdigheter i bruk av PC går ut over både resultatet og den affektive tilstanden hos kandidaten, man mister

troen på seg selv, sier de. Og allerede for mer enn 25 år siden advarte Bill Dunn og David Reay (1989) om at utilstrekkelige ferdigheter med PC og tekstbehandling «forurenser» funn om forbedret tekstlengde og -kvalitet. Harold S. Madsen (1991) kaller det «double jeopardy» når en i vanvare vurderer ferdigheter på PC sammen med språkferdigheter.

Wolfe og Manalo (2004) har en hypotese om at elever med svake PC-ferdigheter må «oversette» fra tanker til tastetrykk, og at elever som skriver på et andrespråk dermed får *to* oversettelsesprosesser. En hypotese om kapasitet i arbeidsminnet sier at prosesser som ikke er automatisert konkurrerer om kapasitet med prosesser som skaper tekst (Dunn & Reay, 1989; Kellogg, 2001). Det er naturligvis bruk av PC og tekstbehandling som ikke er automatisert i dette tilfellet. Når man kobler dette med oversettelseshypotesen hos Wolfe og Manalo (2004), er det ikke unaturlig at tekstene blir svakere hos andrespråksskrivere som skriver på PC. Dette kan underbygges med resultatene for Norge i den internasjonale undersøkelsen PIAAC[5] (Bjørkeng, 2013). Deltakerne her var et utvalg av befolkningen mellom 16 og 55 år. Den viser at eldre og innvandrere skårer signifikant dårligere enn resten av den voksne befolkningen når det gjelder leseferdigheter (literacy) og problemløsning i et IKT-miljø. I slik problemløsning er skåren på ferdighetene tredelt, med nivå 1 som laveste nivå. Bare 24 % av innvandrerne er på nivå 2 eller 3, mot 45 % av de norskfødte (Bjørkeng, 2013). Skriving er ikke undersøkt, men lesing inngår i skriving, og problemløsning i IKT defineres som å ta til seg informasjon, tolke oppgaven og løse den ved hjelp av verktøy på en datamaskin. Disse kravene stilles også i delprøven i skriving på Norskprøven. I denne sammenhengen er «eldre» definert som over 55 år, så det er en litt eldre gruppe enn i min undersøkelse, men allerede fra 44 år synker resultatene på lesing, viser Birgit Bjørkeng (2013). Jeg tror at resultatene fra denne undersøkelsen kan si noe om skriving på PC for voksne innvandrere. Bjørkeng (2013) forklarer innvandreres svakere resultater utelukkende med at oppgavene er på norsk, men dette tror jeg er for enkelt. Det kan tenkes at lese-, skrive- og dataferdighetene også virker inn, jamfør de to hypotesene jeg nevnte tidligere i dette avsnittet (Kellogg, 2001 og Wolfe & Manalo, 2004). Oppgavene i Norskprøven er for øvrig også på norsk.

I en oversikt over studier som sammenligner testing i flere fagområder med penn og PC, viste Hong Wang og Chingwei David Shin (2009) at det er vanskelig å slå fast at erfaring med PC

[5] The Programme for the International Assessment of Adult Competencies, oversatt med Den internasjonale undersøkelsen om lese- og tallforståelse, drevet av OECD.

virker inn på testresultatene, men at de forskjellige fagene har forskjellige resultater. Også måten dataerfaring måles på har en betydning, hevdet de. I 2006 skrev Chapelle og Douglas at vi vet lite om hvordan liten erfaring med PC påvirker resultater i språktesting, men at stadig flere får erfaring med verktøyet etter hvert. De kjente ikke til noen studier som så direkte på dette i testing av andrespråk, og syntes det var merkelig etter to tiår med testing på PC. De anbefaler at kandidatene får støtte til å beherske teknologien før prøver avholdes, slik at variansen i kjennskap til verktøyet i minst mulig grad påvirker variansen i prøveresultatet (Chapelle & Douglas, 2006, s. 17). Jeg sier mer om slik varians i neste underkapittel. I en skrivetest i andrespråkssammenheng er det også viktig at kandidatene får nok tid, slik at de får tenkt gjennom hva de vil gjøre før de begynner å skrive, hevder Liz Hamp-Lyons og Barbara Kroll (1996). Dette er særlig viktig for de som ikke har mye erfaring med tester som har tidsbegrensning, og hva som forventes av dem, fortsetter de. Testens validitet minker hvis tid er en avgjørende faktor for flertallet av kandidatene, eller for spesielle grupper blant kandidatene, sier Sandra Murphy og Kathleen Blake Yancey (2008, s. 371) i en forskningsoversikt. Det virker rimelig å anta at tid blir en viktig faktor hvis PC-kunnskapene er svake og tastaturferdighetene ikke er automatisert.

Hvis vi antar at dette er riktig, blir det interessant å vite hvor mange dette kan gjelde, altså hvor mange eldre som deltar på norskopplæring og norskprøver. I de nye norskprøvene registreres ikke kandidatenes bakgrunnsopplysninger som før[6], men Statistisk Sentralbyrå har tall for antall deltakere i norskopplæring fra 2014. Da var 6239 deltakere mellom 36 og 45 år, 1773 deltakere var mellom 46 og 55 år, og 415 deltakere var over 55 år. I tillegg kommer de som ikke deltok på norskopplæring, og de som ikke er fordelt eller er feilfordelt fordi de ikke har fødselsnummer (SSB, 2016). Avgrensningen her er ikke lik som i mitt prosjekt, men det er rimelig å anta at cirka 5000 deltakere i norskopplæring er 40 år eller mer. Naturligvis tar ikke disse Norskprøven samme år, men hvis vi antar at de tar Norskprøven etter to til fem år, kan det være fra 2500 til 1000 prøvedeltakere årlig som er 40 år eller mer.

I dette avsnittet har jeg vist at å skrive på PC er noe annet enn å skrive med penn og papir, og at det krever tid og trening å mestre dette. Det er grunn til å tro at dette særlig gjelder eldre og innvandrere, som ofte har mindre erfaring med bruk av dette skriveverktøyet. Hypoteser om

[6] Personlig kommunikasjon, e-post fra Cecilie Hamnes Carlsen 6.7.2015.

en ekstra oversettelse for innvandrere og konkurranse om kapasitet i arbeidsminnet når de i tillegg skal skrive på PC kan forklare forskjellene. Jeg antar at mellom 1000 og 2500 personer over 39 år tar Norskprøven hvert år. I det neste avsnittet skal jeg se på hvordan språktesting har utviklet seg, med et særlig fokus på testing av skriving. Deretter vil jeg se på validitet i språktesting, og på konstruktvaliditet og rettferdighet spesielt.

2.5 Språktesting

Språktesting har en svært lang historie. Bernard Spolsky (2008) trekker testingens historie helt tilbake til det kinesiske keiserdømmet for et par tusen år siden, og i England kan historien føres tilbake til 1700-tallet, sier Cecilie Carlsen (2000). Dette kan kalles den førvitenskapelige perioden av språktestingens historie (Spolsky, 1978), der sensorenes dømmekraft var enerådende. På tidlig 1900-tall gikk språktestingen inn i den psykometrisk-strukturalistiske perioden eller retningen hvor prøver skulle rettes objektivt, og man var skeptisk til sensorer, derfor skulle fri produksjon unngås. Først rundt 1960 fikk fagfeltet sine egne teorier, forskning, konferanser og tidsskrifter. Særlig Robert Lados *Language Testing: the Construction and Use of Foreign Language Tests* (1961) ble sentral og regnes som den første fagboken i feltet. Påvirket av strukturalistisk lingvistikk skulle språket som et *system* testes. Det skulle deles opp i sine enkelte bestanddeler som skulle testes separat (McNamara, 2000).

Den psykolingvistiske-sosiolingvistiske perioden eller retningen i språktesting var preget av Noam Chomskys oppgjør med strukturalismen og behaviorismen og mindre fokus på objektive tester (Carlsen, 2000). Tidlig på 1970-tallet begynte den kommunikative perioden med Dell Hymes. I motsetning til den rådende generative lingvistikken, utvidet Hymes i 1972 kompetansebegrepet til også å gjelde språket *i bruk* (Hymes, 1972, s. 282). Inspirert av etnografi så han at språkbrukere også for eksempel tar hensyn til det som er vanlig, passende og mulig i språket. Både kunnskap om språk og kunnskap om språkbruk er sentrale hos Hymes og danner *kommunikativ kompetanse*. Det neste viktige forskningsbidraget sto Michael Canale og Merrill Swain (1980) for. De utviklet en tredelt modell for kommunikativ kompetanse: *Grammatisk kompetanse* som besto av kjennskap til leksikon og regler for morfologi, syntaks, semantikk og fonologi, *sosiolingvistisk kompetanse* bestående av sosiokulturelle regler og regler for diskurs og til slutt *strategisk kompetanse*, det vil si verbale og ikke-verbale strategier som kompenserer for sammenbrudd i kommunikasjonen (Canale & Swain, 1980, s. 29f).

Endelig er vi framme ved forskningsbidraget som i stor grad preger språksynet i Det felles europeiske rammeverket for språk (Udir, 2011), og dermed også Norskprøven for innvandrere (Carlsen, 2000). Det er Lyle F. Bachmans bok *Fundamental considerations in language testing* (1990). Han bygger på forskningen gjort av bl.a. Hymes (1972) og Canale og Swain (1980), utvidet av Canale i 1983. Men Bachman går også videre, og introduserer to nye trekk: Det ene er skillet mellom «kunnskap» («knowledge») og «ferdighet» («skill»), det andre er at han prøver å karakterisere interaksjonsprosessene mellom komponentene i modellen og konteksten for språkbruk (Fulcher & Davidson, 2007, s. 42). Det er særlig Bachmans modell for *communicative language ability* som har fått innflytelse. Begrepet kan oversettes med *kommunikativ språkferdighet* (Pedersen, 2008).

Bachmans språkferdighet er delt i tre: *Språkkompetanse, strategisk kompetanse* og *psykofysiologiske mekanismer.* Dette siste viser til de nevrologiske og psykologiske prosessene i utførelsen av språket som et fysisk fenomen (lyd og lys) (Bachman, 1990, s. 84). Språkkompetansen består av organisatorisk kompetanse og pragmatisk kompetanse. Den organisatoriske kompetansen kan deles i grammatisk kompetanse og tekstuell kompetanse, altså de ferdighetene en språkbruker trenger for å kontrollere formelle strukturer i språket ved å danne og forstå setninger og ordne dem til tekster. Den pragmatiske kompetansen kan deles i illokusjonær kompetanse og sosiolingvistisk kompetanse, det vil si det som har med språkhandlinger, språklig variasjon, register, kulturelle referanser og språklige bilder å gjøre. Strategisk kompetanse er en dynamisk prosess hos språkbrukeren som vurderer relevant informasjon i kontekst og kompenserer eller endrer sine ytringer etter behov. Den strategiske kompetansen påvirkes også av språkbrukerens kunnskapsstrukturer, det vil si kunnskap om verden. Som eksempel på strategisk kompetanse nevner Bachman evnen til å tilpasse beskrivelsen av en reiserute etter hvor den reisende har sitt utgangspunkt (Bachman, 1990, s. 101).

Dette er altså den teoretiske bakgrunnen for utviklingen av Rammeverket (CEFR) og dermed Norskprøven for innvandrere (Carlsen & Moe, 2014, s. 2034). Bachman har fått betydelig kritikk, blant annet fra McNamara (2003), som viser at Bachmans modell primært er psykologisk, mens språkbrukens sosiale kontekst ikke er nevnt. Bachman (og Messick) tar også utgangspunkt i at testutviklerne definerer konstruktet som skal testes, sier McNamara (2003), mens det i virkeligheten er mer pragmatiske og politiske hensyn som ligger til grunn. Senere er Bachmans modell videreutviklet og gjort mer praktisk anvendelig i Bachman og Palmer (1996) og Bachman og Palmer (2010), som jeg kommer tilbake til. I lys av kritikken

av Bachman, utvidet Tim McNamara og Carsten Roever (2006) den sosiale dimensjonen av språktesting.

I tråd med teoriene legger den kommunikative språktesten vekt på *direkte* testmetoder, kandidatens *bruk* av språket, *interaksjon*, for eksempel i samtale, den er kandidatstyrt, og dermed *uforutsigbar*, den bruker *autentiske tekster* og *integrerer* de fire språkferdighetene lese, lytte, snakke og skrive (Carlsen, 2000). Glenn Fulcher hevder også at den bare vurderer om kandidaten oppnår den intenderte kommunikative effekten (Fulcher, 2000). Han er helt på det rene med at dette siste er vanskelig å isolere og spesifisere. Jeg tror likevel at Norskprøven til en viss grad legger mer vekt på kommunikasjon enn korrekthet, for eksempel ved at den bygger på hva kandidaten *kan*, og ved at kandidaten viser at han kjenner til visse trekk, men at disse ikke trenger å være konsekvente i prøven. Marte Monsen (2008) antyder riktignok i sitt mastergradsarbeid at Norskprøve 3 ikke fullt ut måler strategisk kompetanse, på grunn av det som ser ut som et omfattende fokus på formelle ferdigheter.

Så sent som i 1995 sa Anne Golden, Anne Hvenekilde og Else Ryen i forordet til en egen utgave om testing av tidsskriftet NOA at interessen for og kompetansen om testing i Norge var ganske liten (Golden, Hvenekilde, & Ryen, 1995). Senere har den økte bruken av tester i skolen slått inn også i Norge. Vi har fått nasjonale prøver, PISA[7] og andre internasjonale prøver i grunnskolen og Norskprøven for innvandrere, noe som har gitt en økning både i interesse for og kompetanse i språktesting. Universitetet i Bergen er særlig aktive innenfor forskning på testing av voksnes andrespråk, spesielt i forbindelse med Norskprøven og tilknytningen til CEFR. Også Høgskolen i Hedmark og andre forskningsmiljøer har bidratt. Mye av denne forskningen har dreid seg om kravene man kan stille til voksne innvandrere som skal delta i det norske samfunnet (Monsen, 2015). Slik sett føyer mitt prosjekt seg inn i rekken av disse. Jeg vil nå gå over til å beskrive hvordan testing av skriving har blitt gjort.

2.5.1 Testing av skriving

Som i språktesting generelt, var etterkrigstiden i testing av skriving preget av indirekte, objektive tester med fokus på høy reliabilitet, det vil si at to testpersoner som avgir samme svar får samme karakter, eller at to lærere gir samme karakter til samme prøve. Det var vanlig å teste enkeltdeler i språket, for eksempel i form av luketester, flervalgstester eller

[7] PISA er Programme for International Student Assessment, drevet av OECD.

grammatikalitetstester, ikke fri skriveproduksjon. Dette gjaldt i særlig grad i USA. I Storbritannia begynte man allerede på 1940-tallet å fokusere på direkte tester av skriving, og dette fikk etter hvert ringvirkninger også til USA (Hamp-Lyons, 1990). Det var stort fokus på høy reliabilitet, men av det fulgte lav validitet, det vil si at prøven ikke måler noe som er relevant (Fulcher, 2000). En prøve som måler hvilken setning som passer inn i en tekst kan ikke måle hvordan en kandidat skriver brev, for eksempel. Etter hvert ble press fra utdanningsinstitusjoner og foreldre større, og direkte skriveprøver har blitt vanlige, ofte i kombinasjon med andre typer tester.

På 1960-tallet ble det utviklet to måter å vurdere skriveprøver på, analytisk og holistisk. Den analytiske tar utgangspunkt i trekk i teksten som vurderes, det kan være idéer, form, stil, mekanikk og ord. Disse trekkene kan vektes og få poeng, som summeres til en karakter. Holistisk vurdering gir en karakter basert på et sett med kriterier som sensorene har trent på å evaluere (Huot & O'Neill, 2009). Norskprøven bruker holistisk vurdering. Senere er andre vurderingsmåter utviklet. Fra sent på 1970-tallet har det vært vanlig at kandidatene skriver sammenhengende, meningsfulle tekster på målspråket. Det kunne være brev, essay eller andre, mer akademiske sjangre i kommunikativ retning. Dette var inspirert av Robert Kaplans (1966) analyser av retoriske forskjeller mellom kulturer, og at ulike testorganisasjoner bygget på Carrols (1975) modell av de fire språkferdighetene skriving, lesing, muntlig produksjon og lytteferdighet (Cumming, 1997).

Også testmetodisk skjedde det mye på 1970- og 1980-tallet, flere nye metoder ble tatt i bruk (Bachman, 1991). Blant disse er kriteriereferterte prøver, slik som Norskprøven er. Den sier at kandidaten må være på et visst nivå for å få visse rettigheter, for eksempel å begynne på høyere utdanning. Det er ikke relevant hvilke resultater andre kandidater har fått, eller hvor mange som havner på hvert nivå. Kandidaten måles bare mot et nivå på en skala, med visse definerte kriterier. Det blir dermed en sammenheng mellom skåre på prøven og det kandidaten faktisk kan. «…they are appropriate for making absolute decisions», sier Bachman (2004, s. 32) om kriteriereferterte prøver. Da er det naturligvis viktig å utarbeide gode kriterier, og det er her Rammeverket (CEFR) kommer inn. Rammeverket er ikke språkspesifikt (Carlsen, 2012), og det må derfor tilpasses de enkelte språk. Prosjektet Norsk Profil er et første forsøk på å tilpasse de generelle lingvistiske skalaene i Rammeverket til norsk, og bygge på systematiske empiriske studier av språk i bruk (Carlsen, 2012). Det er også viktig med konsensus rundt beskrivelsene i skalaen (Spolsky, 2008), og at sensorer får god opplæring i vurderingsskalaene og har jevnlig oppdatering for at reliabiliteten og validiteten skal bli best mulig (McNamara,

2000; Moe, 2008). Ved Norskprøven som Norsk Språktest er faglig ansvarlig for, har sensorene et vurderingsskjema med formidlingskriterier og språklige kriterier (se Vox, 2016g) som de får opplæring i, og som brukes ved sensurering. Carlsen (2003) finner at dette gir en klar økning i enighet mellom sensorene (intersensor-reliabilitet).

2.5.2 Validitet

I tillegg til teoriene for språktesting, som jeg drøftet i et tidligere avsnitt, har språktestingen også en praktisk målsetning, nemlig å utvikle gode tester i språklige ferdigheter og dessuten å kvalitetssikre eksisterende tester (Carlsen, 2000). Det er under denne siste målsetningen min oppgave kan være et lite bidrag. En viktig del av denne utviklingen og kvalitetssikringen er å vurdere validiteten til prøven. I det foregående har jeg ikke drøftet validitet direkte, men det er grunnleggende i mitt prosjekt. Det har vært (og er fortsatt) vanlig å si at validitet betyr å sikre at prøven måler det den sier at den måler (Halvorsen, 2003), men dette er litt for enkelt innenfor testforskning. «*Validity* is an ominious word», sier Alister Cumming (1996), før han går i gang med å liste opp 16 (!) forskjellige måter ordet validitet har blitt brukt på fra 1930-tallet og framover.

I 1955 la Lee J. Cronbach og Paul E. Meehl et grunnlag for en ny forståelse av validitetsbegrepet og foreslo en måte å bestemme det på. Hamp-Lyons (2003) oppsummerer det tradisjonelle synet på validitetsbegrepet som firedelt. Den enkleste er *face validity*, «…or what looks to an intelligent outsider as if it is valid», sier hun. Hvis en test laget for voksne brukes på barn, kan hvem som helst forstå at den ikke kan brukes. *Content validity* ligner på den foregående, men er begrunnet i faktiske bevis, sier hun videre. Eksemplet hun bruker er at en skrivetest i historie bør be kandidatene skrive om historie. Problemet, sier Hamp-Lyons (2003), er at i skriving er det vanskelig å vite hva kandidatene faktisk gjør, og det er vanskelig å teste skriving uten at det er et innhold. *Criterion validity* er den tredje typen, den viser til et målbart forhold mellom en test og andre mål (kriterier). Dette inneholder *concurrent validity*¸ hvor godt kandidaten gjør det sammenlignet med andre tester på omtrent samme tid, og *predictive validity*, en sammenligning med en annen test i framtida. Problemet med dette er å finne andre kriterier å måle dette mot, ifølge Hamp-Lyons (2003), de fleste vil si at skriving bare kan måles ved at kandidaten skriver. Den siste typen er *construct validity.* Dette viser til en god og riktig beskrivelse av oppførsel i det området som testes, for eksempel er konstruktet i skriving «evne til å skrive». Dette er ganske abstrakt og vanskelig å måle direkte, men testen

må fange opp noen eksempler på god skriving for å ha god konstruktvaliditet etter denne definisjonen (Hamp-Lyons, 2003).

I 1989 kom det som regnes som det viktigste arbeidet innenfor validitet, Messicks (1989) kapittel *Validity* i en forskningsrapport fra ETS, der han var ansatt. Dette er innledningen:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores. As such, validity is an inductive summary of both the existing evidence for and the potential consequences of test interpretation and use. Hence, what is to be valuated is not the test as such, but the inferences derived from test scores -… (Messick, 1989, s. 1).

Messick argumenterer for en integrert vurdering av validitet, der empiri og teori støtter de inferensene og handlingene som foretas på bakgrunn av testskårene. Den som skal konstruere og avholde en test skal altså forutse hvilke konsekvenser som vil komme av testen, og det er ikke testen, men *inferensene* fra testskårene som skal valideres, sier Messick. Litt senere i innledningen står det slik:

> It is important to note that validity is a matter of degree, not all or none. Furthermore, over time, the existing validity evidence becomes enhanced (or contravened) by new findings, and projections of potential social consequences of testing become transformed by evidence of actual consequences and by changing social conditions. Inevitably, then, validity is an evolving property and validation is a continuing process. (Messick, 1989, s. 1).

Her hevder han at det er grader av validitet, og at validiteten kan endres, bli motsagt eller sees i nytt lys. Validitet utvikler seg og er en pågående prosess, sier han. Vi ser at det er en ganske annen definisjon av validitet enn den tradisjonelle. Også hans syn på konstruktvaliditet var nytt, det var nå det overbyggende begrepet som tok andre typer validitet inn i seg, og ble viktigere enn andre typer validitet (Messick, 1989, s. 16). Messicks nye forståelse av validitet ble et de facto paradigme, sier Bachman (2000), selv om mange ikke mener at konsekvensene ikke skal være en del av validiteten (se Messick, 1998, s. 22 for en oversikt). Messick drøfter flere typer validitet, jeg skal ikke gå inn på alle her. En type validitet, konsekvensvaliditet, skal jeg likevel bruke litt plass på. Som vi så i det første sitatet fra Messicks arbeid, er potensielle konsekvenser av testtolkning og -bruk en del av validiteten. Han hevder at feil fordeling i testskåre som skyldes kjønn og etnisk bakgrunn er eksempler på testugyldighet (test invalidity) (Messick, 1989, s. 153). Det er den som er ansvarlig for testen som skal forutse slike problemer og vurdere hvilken effekt de har på validiteten (McNamara, 2000, s. 54). Slik ser vi at Messicks forståelse av validitet blir en kobling mellom etikk og validitet (Bachman,

2000). Glenn Fulcher og Fred Davidson (2007, s. 143) nærmer seg filosofien når de viser at konsekvensene er i høysetet: «As such, we may define any test as its consequences».

Messick hevder at alle typer trusler mot konstruktvaliditeten passer i to kategorier (Messick, 1989, s. 44). Den første kaller han *konstrukt-underrepresentasjon* og den oppstår når testen er for smal og ikke inkluderer viktige dimensjoner ved konstruktet. Testen krever for lite av kandidaten, og kan dermed ikke anses å representere konstruktet (Messick, 1989, s. 45). Den andre typen trussel mot konstruktvaliditeten kaller han *konstrukt-irrelevant varians*. Han sier det slik: «… - the test contains excess reliable variance that is irrelevant to the interpreted construct» (Messick, 1989, s. 45). Det er altså kommet noe inn i testen som ikke tilhører konstruktet, en «forurensning». Messick hevder videre at det er to typer av denne variansen: *construct-irrelevant difficulty* som betegner at noe kommer inn som et tillegg og gjør testen vanskeligere. Hans eksempel på dette er at en fagtest (test av subject-matter knowledge) krever urimelig mye leseforståelse. Dette gir vanligvis lavere skåre for de personene som utsettes for det. Den andre typen varians kaller han *construct-irrelevant easiness*, at noe, for eksempel i testen, gjør at kandidaten får bedre resultat enn uten denne variansen. Messick gir et eksempel: noen elever kan forstå at det lengste svaret alltid er rett i multiple choice-oppgaver. Disse elevene vil tjene på dette, og prøven inneholder da *construct-irrelevant easiness* (Messick, 1989, s. 45). Rent intuitivt virker dette også ganske opplagt. Messick flytter det som tradisjonelt har vært ansett som trusler mot reliabiliten til trusler mot validiteten (Carlsen, 2005). Slik får validiteten en mer sentral rolle, som vi har sett.

Bachman tok opp Messicks arbeid i boka «Fundamental considerations in language testing» i 1990. Der viser han til tre faktorer som påvirker testskåren (Bachman, 1990, s. 165). Dette er kilder til konstrukt-irrelevant varians, sier Carlsen (2004). Hun nevner *faktorer ved prøven* (vanskegrad, oppgavetyper, sensorer, vurderingsskala), *faktorer ved studentene* (kjønn, alder, kulturell bakgrunn, faktakunnskaper) og *tilfeldige faktorer* (studentenes dagsform, omgivelsene for prøven) (Carlsen, 2004). Vi ser at det potensielt kan være problemer når en skriveprøve går fra å utføres med penn og papir til å utføres med PC (Wolfe & Manalo, 2005). Det er faktorer ved prøven og faktorer ved studentene som bør vurderes i dette tilfellet, tilfeldige faktorer er antagelig lik for de to modusene, hvis vi ser bort fra tekniske problemer ved datautstyr, lysforhold på dataskjerm og så videre. Jeg vil drøfte dette videre i oppgaven. Men hvor blir det så av reliabiliteten og de tradisjonelle definisjonene av validitet? Jo, de er fortsatt viktige, men en god test krever mer enn dette, sier Hamp-Lyons (2003, s. 165).

Messicks definisjon av validitet er utfordrende for alle som konstruerer og er ansvarlige for en test, de skal altså forutse mulige tolkninger og misbruk av testskåre, blant annet. For å tette dette gapet mellom teori og praksis i språktesting har Lyle F. Bachman og Adrian S. Palmer (1996) utviklet seks kriterier som bør undersøkes av de som skal evaluere en test, sier Chapelle (2012). Bachman og Palmer (1996) formulerer dette som *testbrukbarhet (test usefulness),* en overbygning som består av reliabilitet, konstruktvaliditet, autentisitet, interaktivitet, innvirkning og praktisk bruk (Bachman & Palmer, 1996, s. 17). Begrepet *testbrukbarhet* har likevel ikke blitt mye brukt i litteraturen senere, sier Fulcher og Davidson (2007), og antar at det skyldes at «degraderingen» av konsekvensvaliditet til testbrukbarhet ikke har vært en utfordring for fagfeltet. Jeg vil se på ett begrep fra Bachman og Palmer (1996). *Innvirkning (impact)* vil si den påvirkningen testen og resultatene har for alle som tar testen eller bruker den senere, også de som tar avgjørelser på bakgrunn av den. For testtakerne er avgjørelser som tas om dem på bakgrunn av testskåre en sentral del av påvirkningen. Bachman og Palmer spør om det er riktig å ta avgjørelser som kan påvirke personers liv på bakgrunn av skåre på en test (Bachman & Palmer, 1996, s. 33). Et sentralt spørsmål i evalueringen er «To what extent can we justify these interpretations?», sier Bachman og Palmer (1996, s. 21). Vi må føre bevis eller rettferdiggjøre «…that the testscore reflects the area(s) of language ability we want to measure, and very little else», sier de videre. Hvis vi sier at den språklige ferdigheten eller konstruktet er *å skrive*, kan vi altså ikke legge særlig mer inn i det. Da må vi i tilfelle kalle ferdigheten noe annet.

I sin bok fra 2010 går Bachman og Palmer et steg videre i å ansvarliggjøre de som utvikler og arrangerer tester, idet de innfører begrepet *ansvarlighet (accountability).* Fordi involverte personer og institusjoner vil påvirkes i stor grad ved bruk av tester, er det opp til de som utvikler og arrangerer tester å demonstrere at den intenderte bruken av testen kan rettferdiggjøres, det gjelder konsekvenser og avgjørelser som er tatt (Bachman & Palmer, 2010, s. 93). Uten dette, sier de, kan testene brukes eller misbrukes av andre senere. Konstruktvalidering blir viktigere etter hvert som stadig flere instanser krever språkprøver (Golden & Monsen, 2015, s. 207). McNamara (2003, s. 467) sammenligner valideringsarbeidet med en rettssak: inferensene til en test jamføres med bevis og motbevis som vektes for eller mot testens brukbarhet.

Når en skal teste skriving på PC, er det viktig å vurdere om kandidatene er kjent med å skrive på PC, og hvilken effekt PC-ferdigheter har på skriveferdighetene på en test, sier Weigle (2002, s. 105). Hun sier videre om tastaturferdigheter:

> Unless such skills are part of the construct (as they might be, for example, in a test for office workers) it is clearly inequitable to require students with weak or non-existent keyboarding skills to use a computer rather than pen and paper on writing tests» (Weigle, 2002, s. 237).

Vi ser at dette er i tråd med tankene til Bachman og Palmer (1996), som referert over. En må altså være tydelig på at det er *skriving på PC* som testes, ikke skriving som sådan. Slik validering er Chapelle og Douglas (2006, 42) også inne på, de hevder at hvis tankeprosessen endres ved at testen er på PC, endres validiteten til inferensene vi kan gjøre. Validiteten er altså endret ved at testen er på PC. Også i veilederen *ETS Standards for Quality and Fairness* er det et krav at validiteten vurderes på nytt når relevante faktorer i testen endres, og teknologi nevnes som et eksempel. Om nødvendig skal en samle inn nye bevis for at testen er valid (ETS, 2014, s. 18). Oppsummert er det enkelt å være enig med Alan Davies (2014, s. 11) i at fra tidlig 1960-tall og fram til i dag har spørsmålene gått fra «how?» og «what?» til «why?» og «should we?».

2.5.3 Rettferdighet

Jeg nevnte i innledningen at Norskprøven for mange elever kan være en «high-stakes test», en viktig prøve som kan forandre folks liv og være vanskelig å gjøre på nytt (Bachman, 2004, s. 12). Og slike prøver er laget for å klassifisere eller velge ut individer, sier Fulcher (2014, s. 1554). I slike livsavgjørende tester og ellers i samfunnet ønsker vi at individer og grupper skal behandles rettferdig. Men hva er rettferdig? Messick (1998, s. 12) sier at konstruktvaliditet og rettferdighet er uløselig koblet til hverandre, og fortsetter:

> Fairness, like validity, is not just a psychometric issue. It is a social value, and there are alternative points of views about its essential features. In essence, fairness implies impartiality, with an absence of prejudice or favoritism. In regard to test use, this impartiality derives from comparable construct validity. A fair test is one that yields comparably valid scores from person to person, group to group, and setting to setting… (Messick, 1998, s. 13).

Rettferdighet er en sosial verdi, den skal være upartisk og ikke forutinntatt. Denne upartiskheten stammer fra sammenlignbar konstruktvaliditet for alle grupper og personer, sier Messick. Likevel er det ikke sånn at alle personer skal få lik skåre, men forskjellig innlæring og utvikling skal dokumenteres i testen, hevder Messick (1998, s. 2). Rettferdighet må gjennomsyre utviklingen av testen, men særlig er valg av konstrukt kritisk fordi det påvirker innhold, format og spesifikasjoner i testen. Den som lager testen må undersøke om alle grupper har like stor mulighet til å få samme skåre, fortsetter han (Messick, 1998, s. 18). Det er viktig

å være klar over at all testing er verdiladet, og våre valg av hvordan tester brukes avslører våre politiske og filosofiske preferanser, sier Fulcher (2014, s. 1554).

Det er viktig å være tydelig på hva konstruktet er og ha et system for hvordan det skal kommuniseres til alle som har med testen å gjøre. Hamp-Lyons (1991, s. 61) sier at en kandidat bør få vite testens hensikt (om det er utestengelse, plassering eller diagnose), format, lengde, antall spørsmål og eventuell vekting, *hva slags skriving som vurderes* (min utheving), kriterier, metode for å beregne skåre, skårens reliabilitet og sensorenes kvalifikasjoner. Dette bør presenteres proaktivt, sier hun, altså at kandidaten ikke skal måtte spørre om informasjon. Elever i voksenopplæring får sannsynligvis vite det meste av dette fra sin lærer, men Vox har et ansvar for å kommunisere dette til prøvesteder, skoler og lærere og til kandidater som melder seg på test uten å gå på kurs. Den beste måten å nærme seg idealet om rettferdighet på er å holde konstruktirrelevant varians så lav som mulig, står det i veilederen fra ETS, som jeg nevnte over. Det er umulig å undersøke om prøven er rettferdig for alle grupper, men man skal gjøre det for grupper man har grunn til å tro blir påvirket (ETS, 2014, s. 19). Da TOEFL-testen[8] ble endret fra at papirbasert til PC-basert, gjorde ETS en undersøkelse for å kartlegge kandidatenes ferdigheter på PC før endringen ble iverksatt (Kunnan, 2005, s. 788). Bachman og Palmer (2010, s. 40) anbefaler at man ser på noen personlige attributter som ikke er en del av språkferdigheten (som de nå kaller *language ability*), men som likevel kan påvirke prestasjonen på en språktest. Det er *alder, kjønn, nasjonalitet, oppholdsstatus, lengde på opphold, morsmål, nivå og lengde på utdanning* og *type og mengde av forberedelse eller tidligere erfaring med den gitte testen* (Bachman & Palmer, 2010, s. 41). Noen av disse er viktige elementer i vurderingen av språkprøven og av mine resultater, særlig den første og den siste.

Vox er medlem i *Association of Language Testers in Europe* (ALTE) (ALTE, 2016). I deres Praksiskodeks ser vi at ALTE-medlemmene skal utvikle prøver, tolke prøveresultat, arbeide for rettferdighet og informere kandidater (ALTE, 2007, s. 2). De skal også «Vedta prosedyrer som skal sikre at resultatene på prøven først og fremst skyldes de ferdighetene prøven skal måle og ikke irrelevante faktorer som for eksempel kjønn eller etnisk bakgrunn» (ALTE, 2007, s. 4). Vi ser at det er validitets- og rettferdighetsargumentene som skinner gjennom. Ansvaret

[8] TOEFL er Test of English as a Foreign Language, drevet av ETS.

for å vedta disse prosedyrene hviler på Vox i Norge. I Norskprøven kan man gå ut fra at elever med lite trening på PC vil bli påvirket av at testen blir gjennomført på PC.

2.6 Oppsummering av teorier og tidligere forskning

Jeg har vist at andespråksbrukere har et mellomspråk med sin egen grammatikk. Jeg nevnte også at skriveprosessen i førstespråket kan deles i tre: planlegging, transkribering og revisjon. Teorier om skriveprosessen i andrespråket er kompliserte og mangelfulle og har fått liten betydning i forskningen. Det er sannsynlig at den kognitive belastningen ved å skrive på et andrespråk er høyere enn i førstespråket, og mye av tiden går med til formulering tidlig i andrespråksinnlæringen. Andrespråksskriving er mer krevende enn førstespråksskriving, og det resulterer i at produktet generelt blir svakere hos elever som ikke skriver på sitt morsmål. Å skrive på PC gir en del fordeler framfor å skrive med penn og papir hvis elevene får nok og relevant undervisning eller har tilstrekkelige ferdigheter med tekstbehandling. Dette siste gjelder ikke nødvendigvis eldre innvandrere, som i varierende grad har erfaring med bruk av PC. De vil dermed få en «dobbelt oversettelse».

Språktesting har en lang historie, men moderne språktesting startet på 1960-tallet. Et sosialt syn på språktesting har sitt utgangspunkt i teorier fra Messick (1989) og Bachman (1990). Der er konstruktvaliditet et sentralt begrep, nært knyttet til rettferdighet. Det påhviler den som utvikler en test å undersøke om spesielle grupper kan få systematisk dårligere resultater ved at prøven innføres. Jeg skal se om skriveredskapet har en betydning for kvaliteten i tekstene for andrespråkselever, og særlig se på eldre kandidater. I det neste kapittelet vil jeg gjøre rede for metoden jeg har brukt, derunder testkonstruksjon, deltakere og gjennomføring av testen, måleenheter og statistiske metoder jeg har brukt og valg jeg har gjort.

3. Metode

Fordi jeg ikke fikk tilgang til autentiske tekster fra Norskprøven har jeg benyttet et kvasieksperimentelt forskningsdesign, der to grupper av elever med så lik bakgrunn som mulig fikk en skriveprøve i hver sin modus. Disse prøvene har jeg transkribert og korrigert før jeg testet dem kvantitativt med statistiske metoder, deskriptivt og inferensielt. I dette kapittelet viser jeg først hvordan testen ble bygget opp så nær opptil Norskprøven som mulig, deretter gjør jeg rede for elevenes bakgrunn, før jeg viser hvilke måleenheter for kvalitet i skriving jeg valgte. Kapittelet avsluttes med en diskusjon om statistiske metoder og valg jeg har gjort.

3.1 Testkonstruksjon

For å kunne si noe om hvordan voksne elever skriver på Norskprøven, var det hensiktsmessig for meg å legge min prøve så nær opptil Norskprøven fra Norsk Språktest som mulig, slik den gjennomføres to ganger i året. Prøvene gjøres ikke tilgjengelige etter at de er avholdt, men det ligger en eksempelprøve på nettstedet til Vox (Vox, 2016a). I innledningen sa jeg litt om hvordan Norskprøven er bygget opp, men jeg skal utdype mer her. Norskprøven er identisk for A2- og B1-nivå. Prøven i delferdigheten *skriftlig produksjon* varer i 90 minutter. Den inneholder tre oppgaver med ulik vanskegrad, og kandidatene må svare på alle tre for å få en vurdering. Oppgave 1 består i å beskrive et bilde, og kandidatene anbefales å skrive mellom 80 og 100 ord. Oppgaveordlyden er: «Beskriv bildet. Skriv mellom 80 og 100 ord». Her er det for eksempel et bilde fra en by hvor det foregår mye, eller fra en familiesituasjon med aktiviteter i huset som skal beskrives. Kandidatene kan bruke et enkelt ordforråd, og oppgaven innbyr til enkle setninger av typen «Jeg ser…». Dette er en kjent oppgavetype for elever som har hatt begynnerundervisning, og den finnes også i A1/A2-prøven og måler dermed på A2-nivå (Vox, 2016c).

For at min prøve (vedlegg 1) skulle være ukjent for alle, laget jeg et gatebilde i tegneverktøyet Creaza (2016) til oppgave 1, der en del personer, aktiviteter og gjenstander skulle beskrives. Ordlyden i min oppgave 1 var identisk med ordlyden hos Norsk Språktest. Oppgave 2 var en narrativ oppgave. I eksempelprøven er instruksjonen til kandidaten slik: «Skriv en e-post til en venn. Skriv mellom 80 og 200 ord». Under denne er det knapper som skal illudere et e-post-program, men de virker ikke, og kandidaten skal bare skrive tekst i et åpent felt og trykke «Neste». Igjen bygger oppgaven på kjent ordforråd, dette har elever i begynnerundervisning

gjort før, og de kan velge kjent syntaks og unngå vanskelige elementer. Denne oppgaven finnes også i A1/A2-prøven, og den måler på A2-nivå i begge prøvene (Vox, 2016c). Jeg ville unngå at elevene kopierte eksempelprøven, men ville at de skulle ha noe kjent, så i tillegg til eksempelprøvens oppgaveordlyd, la jeg til «Fortell hva du gjorde sist sommer». Dette har antagelig alle elever gjort før, og jeg gjennomførte prøven om høsten, så det burde være ferskt i minne. I oppgave 3 skal kandidaten uttrykke egne meninger. Ordlyden i eksempelprøven er slik: «Skriv en e-post til busselskapet. Skriv minimum 80 ord. Busselskapet vil legge ned noen av bussrutene der du bor. Dette skaper problemer for deg og andre i nærmiljøet. Skriv en e-post til busselskapet og forklar hvorfor det er viktig at bussene går like ofte som før» (Vox, 2016a). Igjen likner dette et e-post-program. Vi ser på ordlydens lengde, ordforråd, syntaks og forventet talehandling at dette er et nivå over de andre oppgavene. Likevel var ordforrådet slik at det burde være kjent, og talehandlingen «Uttrykke egne meninger» bør også være kjent fordi det er en del av læreplanen og lærebøkene.

Ord som elevene har behov for i livssituasjonen sin er ofte kjent, selv om de er komplekse eller lavfrekvente, sier Golden (2014, s. 134). En kan forvente at «busselskap» og «nærmiljø» kan være slike ord i eksempelprøven. I Norskprøven er denne oppgaven også med i B1/B2-prøven, den måler da på B1-nivå (Vox, 2016c). Jeg valgte også et antatt kjent domene, skolen: «Skriv en e-post til skolen. Skriv minimum 80 ord. Skolen skal flyttes til en by lengre fra deg. Dette skaper problemer for deg og flere av vennene dine. Skriv en e-post til skolen og forklar hvorfor det er viktig at skolen ikke skal flyttes.» Ordforrådet i min prøve var kanskje noe enklere enn i eksempeloppgaven, men syntaksen var temmelig lik. Layout for min prøve var helt lik prøven fra Norsk Språktest, og både i den, min digitale prøve og den papirbaserte prøven var det mulig å bla fram og tilbake. En kunne gjøre oppgavene i den rekkefølgen en ønsket. Papirversjonen av prøven var helt lik den digitale versjonen, men det var linjerte områder mellom oppgavene med plass til å skrive på.

Norskprøven er bygget opp med økende vanskelighetsgrad. Macaro (2003, s. 224f) hevder, med støtte fra forskning, at det er forskjell på hvor vanskelige ulike sjangre i oppgaver er for nybegynnere og litt erfarne andrespråksskrivere. Enklest er beskrivende oppgaver, deretter kommer narrative oppgaver, mens resonnerende (expository) oppgaver er vanskeligst. Slik er også Norskprøven bygget opp. Argumenterende oppgaver kommer først inn på B2-nivå. Macaro forklarer forskjellene i vanskegrad med ulike krav til ordforråd, syntaks og idiomatiske uttrykk, men jeg tror også ulike retoriske forventninger i sjangrene og ulike talehandlinger har en betydning. Oppgave 1 var relativt enkel (Macaro, 2003), og oppfordret

indirekte kandidatene til å skrive setninger av typen «Jeg ser…», som nevnt. Dette gir et enkelt språk, og tekstene herfra egner seg ikke til den typen undersøkelse jeg gjør her. Elevene skriver mye rart til bilder, ifølge Anne Golden[9]. På oppgave 3 i min prøve var det sannsynligvis noen som fikk litt lite tid, noen av tekstene preges av hastverksarbeid og en brå avslutning. Likevel fyller de kravene til tekstlengde for B1 på minimum 80 ord, men ikke kravene i mine variabler på 85 ord (OVF) og 100 ord (vocd D og MTLD). Variablene forklarer jeg mer om senere i kapittelet. Av disse årsakene vil jeg bare bruke tekstene skrevet til oppgave 2 som data for de fleste variablene i prosjektet. Jeg går mer inn på lengden på besvarelsene senere. For variabelen *tekstlengde* (TL), som summerer alle tekstene kandidaten har skrevet, vil også oppgave 1 og 3 være inkludert. Jeg skal nå gå over til å beskrive deltakerne og hvordan testen ble gjennomført. De statistiske metodene og valgene gjør jeg rede for senere i kapittelet.

3.2 Deltakere og gjennomføring av testen

I dette avsnittet bruker jeg en del statistiske metoder som jeg gjør nærmere rede for i et eget avsnitt senere i oppgaven, kapittel 3.4. I forskning ønsker en ofte å teste hypoteser om noe en har observert. Det optimale er gjerne å benytte et eksperimentelt forskningsdesign, men det kan være vanskelig å få til i en skolekontekst. I stedet må man ofte benytte den skolen en jobber på og de elevene en selv har. Likevel er dette antagelig mer representativt for de faktiske forholdene i skolen enn det eksperimentelle designet, sier Herbert W. Seliger og Elena Shohamy (1989, s. 148). Når forskningsdesignet bygger på en situasjon som allerede eksisterer, for eksempel at elevene ikke er tilfeldig valgt ut, kalles designet *kvasieksperimentelt,* fortsetter de. Alt annet enn utvalget kan være identisk med et rent eksperimentelt design. Med to uavhengige grupper er det viktig å få den usystematiske variasjonen mellom gruppene så liten som mulig. Hvis dette er godt gjennomtenkt, kan det gi like gode resultater som et rent eksperimentelt design, sier Zoltan Dörnyei (2007, s. 117), og peker særlig på to viktige måter å forbedre et kvasieksperimentelt design på: Elevene plasseres i grupper av forskeren, og førtest-forskjellene må være minst mulige. Dette siste deler Dörnyei inn i to igjen: Den første er å matche elever i de to gruppene basert på én eller flere bakgrunnsvariabler, for eksempel at to jevngamle jenter med omtrent like lang utdanning kommer i hver sin gruppe. Den andre er å gjøre en analyse av kovarianse (ANCOVA) etter

[9] Personlig opplysning 30.10.2015 ved et masterseminar på Høgskolen i Hedmark, campus Hamar.

skriveprøven for å justere for førtest-forskjellene (Dörnyei, 2007, s. 117f). Jeg har gjort det som Dörnyei nevner her, unntatt å gjøre en ANCOVA. Det vil føre for langt å gjøre dette for alle de ti avhengige variablene.

Det hadde også vært en mulighet å bruke de samme elevene to ganger, altså en parvis test, hvor de skriver én gang for hånd og én gang med PC. Da unngår man førtest-forskjeller, og får med det større styrke statistisk. Likevel får man to andre ulemper: For det første *treningseffekt*, at deltakerne opptrer annerledes andre gang enn første gang testen gjøres. De kjenner situasjonen og/eller oppgavene, og gjør ikke det samme som om de var ukjent med dette. For det andre er det *kjedsomhet*, at deltakerne er slitne eller leie fordi de har gjort oppgaven før (Field, 2000, s. 211). Dette siste kan kanskje delvis unngås ved at det er noe tid mellom de to testene, men da kan elevene ha lært noe nytt i mellomtiden. Field (2000) viser at de to effektene også kan unngås (med et såkalt ABBA-design), men for meg var det også praktiske årsaker til at jeg ville unngå denne metoden. Det er vanskelig å be voksne mennesker som kanskje betaler for hver time om å jobbe med noe som er såpass krevende to ganger uten at de føler at de får nok igjen for det. Hvis noen er lite motivert, vil de ikke gjøre sitt beste, som om de var på en reell test, eller de vil utebli fra testen. Dermed valgte jeg et én-prøve-design.

Den uavhengige variabelen *skrivemodus* gir to grupper som skal være mest mulig balanserte. I prioritert rekkefølge forsøkte jeg å få balanse i alder, kjønn, utdanning, botid i Norge og antall norsktimer. Flere av disse er usikre og baserer seg delvis på elevenes rapportering. Utdanning ble angitt i antall år, men det kan variere hva elevene definerer som et år, noen kan ha hatt korte skoleuker noen måneder i året, mens andre har lengre skoleår enn i Norge. Type utdanning kan også ha betydning, en kan anta at en praktisk yrkesutdanning har mindre skriving enn en teoretisk utdanning, og at høyere utdanning har mer skriving enn grunnutdanning. Botid i Norge er ment fra den måneden de kom til Norge målt i måneder, men enkelte husker feil, misforstår, eller har vært i andre land i mellomtiden. Antall norsktimer er også beheftet med feil: timene føres i et statlig rapporteringssystem, men det er stadig feilføringer, og hvis familiegjenforente gifter seg etter at de kommer til Norge, telles timene

først etter at ekteskapet er inngått. Heller ikke norsktimer i utlandet eller utenfor det offisielle systemet NIR[10] registreres.

Tre klasser ved en skole på Østlandet deltok, de ble fordelt på én gruppe som skrev med penn på papir og én gruppe som skrev med PC i Microsoft Word. Klassene inneholdt stort sett elever på A2/B1-nivå i skriving, men lærerne rapporterte at noen var under A2-nivå, og noen kunne være i øvre sjikt av B1-nivå. Totalt deltok 47 elever på testen, de to gruppene satt i hvert sitt rom. Jeg var prøveleder i PC-gruppa, mens en annen lærer var prøveleder i penn og papir-gruppa. Alle som ville ha fikk utdelt papir til planlegging og notater. To elever hadde ikke besvart alle tre oppgavene, og deres besvarelser ble derfor fjernet fra analysen. Det viste seg at tre elever i penn og papir-gruppa og fire elever i PC-gruppa ikke hadde nådd A2-nivå, disse svarene ble også fjernet. (Se vurderingsskjema, Vox, 2016g). De hadde fått en for vanskelig oppgave i forhold til sine ferdigheter, og det var ikke relevant å ha deres besvarelser med videre. Det ble 19 elever i hver gruppe igjen, til sammen 38 elever. Disse fordelte seg slik fra de tre klassene: seks, fem og åtte elever i penn og papir-gruppa, og seks, sju og seks elever i PC-gruppa. Elevene besvarte tre oppgaver på 90 minutter, som i Norskprøven. På Norskprøven er det ikke retteprogram, derfor ble denne funksjonen i Word skrudd av for PC-gruppa etter at oppgaveteksten var lagret på hver enkelt PC. Ingen hadde tilgang til internett. Besvarelsene ble lagret på min minnepenn da elevene var ferdige og overført til min PC. Deretter ble de slettet på elevenes PC og på minnepennen.

3.2.1 Alder og kjønn

I begge gruppene var den yngste eleven 18 år, mens den eldste var 52 år. Middelverdien for penn og papir-gruppa var 29,42 år (sd ≈ 9,91), mens det var 29,79 år (sd ≈ 10,99) for PC-gruppa[11]. En Levenes test (Levene, 1960) viste at variansen var homogen (F = 0,923, p = 0,343), og premisser for normalitet var oppfylt. En t-test viste at de to utvalgene kunne komme fra samme populasjon ($t(36)$ = 0,109, p = 0,914). Effektstørrelsen var 0,04 (Cohens d), altså nesten ingen forskjell. Vi ser at gjennomsnittet var høyere og spredningen i alder var litt større for PC-gruppa, men at gruppene var så å si like. Det var 19 elever i hver gruppe. I penn og

[10] Nasjonalt introduksjonsregister, drevet av Integrerings- og mangfoldsdirektoratet (IMDI).

[11] Eksakt fødselsår er brukt. Det er vanlig å legge til et halvt år i statistiske beregninger, fordi det blir feil å anta at alle er født 1. januar (Kristiansen, 2010). I denne sammenhengen er dette uten betydning.

papir-gruppa var det 10 kvinner og 9 menn, i PC-gruppa var det 9 kvinner og 10 menn. Vi kan anta at alder og kjønn ikke har noen betydning for undersøkelsen.

3.2.2 Utdanning

Elevenes utdanning varierte mellom 3 år og 16 år for penn og papir-gruppa, og mellom 2 år og 16 år for PC-gruppa. Gruppene var ikke normalfordelt, de hadde en bimodal fordeling. En ikke-parametrisk Levenes test (Nordstokke, Zumbo, Cairns, & Saklofske, 2011) viste at variansen er homogen (p = 0,556), og en Mann-Whitney U-test viste at utdanningen for de to gruppene var ganske lik. Medianen for penn og papir-gruppa var på 12 år, mens medianen for PC-gruppa var på 9 år. Forskjellen mellom de to var ikke statistisk signifikant, U = 160,5, *p* = 0,566. Effektstørrelse: Cohens *r* = 0,095 (Fritz, Morris, & Richler, 2012, s. 12)[12]. Vi ser at penn og papir-gruppa hadde litt lengre utdanning, men ikke i en slik grad at det skulle påvirke undersøkelsen.

3.2.3 Botid i Norge

Botiden varierte fra 14 til 84 måneder for penn og papir-gruppa, og fra 11 til 83 måneder for PC-gruppa. Penn og papir-gruppa hadde to høye verdier på 69 måneder og 84 måneder, PC-gruppa hadde én høy verdi på 71 måneder og én utligger på 83 måneder. Penn og papir-gruppa var i tillegg høyreskjev. En ikke-parametrisk Levenes test viste at variansen var homogen (*p* = 0,796), og en Mann-Whitney U-test viste at botiden for de to gruppene var ganske lik. Median for penn og papir-gruppa var 35 måneder, for PC-gruppa 36 måneder. Forskjellen mellom de to var ikke statistisk signifikant, U = 179,5, *p* = 0,977. Effektstørrelse: Cohens *r* = 0,0047, altså ingen effekt. PC-gruppa hadde litt lengre botid, og litt større spredning i observasjonene, men ikke slik at det skulle ha betydning for undersøkelsen.

3.2.4 Antall norsktimer

Antall norsktimer varierte fra 400 timer til 2060 timer i penn og papir-gruppa og fra 420 timer til 2120 timer i PC-gruppa. Timetallet er avrundet til nærmeste tier. Middelverdien for penn og papir-gruppa var 1148,9 timer (sd ≈ 523,49) og for PC-gruppa 1071,58 (sd ≈ 507,74). En Levenes test viste at variansen var homogen (*F* = 0,028, p = 0,868) og en t-test viste at utvalget

[12] Jeg gir en begrunnelse for å bruke denne utregningen i kapittel 3.4.3.

kunne komme fra samme populasjon ($t(36) = 0{,}462$, $p = 0{,}647$). Effektstørrelsen var 0,15 (Cohens *d*), så det var en ganske liten forskjell. Penn og papir-gruppa hadde litt flere norsktimer og litt større spredning. Likevel kan vi anta at antall norsktimer ikke skulle ha en betydning for undersøkelsen.

3.2.5 Opprinnelsesland

I den norske delen av den internasjonale ALL-undersøkelsen[13] fant Egil Gabrielsen og Bengt Oscar Lagerstrøm (2007) signifikante forskjeller i basisferdigheter (literacy og tallforståelse) mellom det de kalte *vestlige innvandrere* og *ikke-vestlige innvandrere*. Vestlige innvandrere hadde opprinnelse i Vest-Europa, USA, Canada, Australia og New Zealand, mens ikke-vestlige innvandrere hadde opprinnelse i den gamle østblokken og resten av verden.[14] De begrunnet forskjellene med kulturell avstand. Det er grunn til å tro at slike forskjeller også vil påvirke innholdet i skriftlige tekster. Ingen av mine elever hadde vestlig bakgrunn etter denne definisjonen, så jeg har valgt å se bort fra denne variabelen. Jeg kunne i stedet sammenlignet områder i verden med hverandre, men det havnet utenfor denne oppgavens siktepunkt. Alle deltakerne i undersøkelsen er født utenfor Norge av utenlandske foreldre og er selv innvandrere til Norge.

3.2.6 Transkripsjon og korreksjon

Etter prøven skrev jeg de håndskrevne tekstene digitalt. Det ble ikke gjort noen endringer. Fordi to av målene for kvalitet måler på ukorrigerte tekster, ble de tatt på tekstene, både de håndskrevne og de digitale, slik de framsto før korreksjon. Det var *gjennomsnittlig t-enhetslengde* og *antall feil per t-enhet*. Disse skal jeg drøfte i neste avsnitt. I tre tilfeller var jeg usikker på hvilken bokstav som var skrevet i håndtekstene, men etter å ha sammenlignet med andre bokstaver eleven hadde produsert, mener jeg at det ble riktig. Deretter ble alle tekstene korrigert for feil (se vedlegg 2) og de andre målene ble tatt. For noen variabler måtte norske tegn erstattes med andre, jeg sier mer om dette i neste avsnitt. Se vedlegg 2 for hvordan jeg rettet tekstene.

[13] Drevet av OECD og Statistics Canada

[14] Avgrensingen mellom innvandrergrupper brukes også i finansiering av norskopplæringen (IMDI, 2011, s. 40).

Andrespråksbrukeren identifiseres og viser oss sitt språk gjennom feil, sier Jon Erik Hagen (2005). Det er ofte slik at feil er til hinder for forståelsen, og iallfall er det ofte slik at kommunikasjonen ikke flyter godt med mye feil. Språktesting har et normativt utgangspunkt, en sammenligning med en målspråksnorm, sier Trinelise Eriksson og Cecilie Carlsen (2012). Dermed har man også tatt et standpunkt i diskusjonen mellom *akseptabilitet*, som er subjektiv og avhengig av stilnivå, og *grammatikalitet*, i favør av grammatikaliteten, altså brudd på regler (Ellis & Barkhuizen, 2005, s. 56). Ellis og Barkhuizen anbefaler at man bruker grammatikalitet, nettopp fordi den ikke er så subjektiv. Da må man også tenke på forskjellen mellom *åpne* og *skjulte feil*, sier Eriksson og Carlsen (2012) videre. Åpne feil er de som avdekkes ved å lese setningen eller ytringen hvor den opptrer, mens skjulte feil bare kan finnes ved å se på større deler av diskursen. Jeg vil anta at de her mener situasjonskontekst og kulturkontekst. En leser skal ikke måtte resonnere seg fram til hva som er ment (Berggreen & Tenfjord, 2007, s. 34), men på den andre siden må teksten henge sammen med konteksten. Det medfører at man kan definere feil som Paul Lennon (1991, s. 182) gjør: «a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native counterpart». Hvis en ytring ikke er slik en målspråksbruker i samme situasjon ville ha produsert den, har den altså feil, ifølge denne definisjonen. Og det er nettopp sammenlikning med målspråket som er hensikten med å måle nøyaktighet, hevder Kate Wolfe-Quintero, Shunji Inagaki, & Hae-Young Kim (1998, s. 33).

På tross av dialekt- og registerforhold, er det en viss konsensus blant språkgrupper for hva en feil er, sier Charlene Polio og Mark C. Shea (2014, s. 10). Jeg vil praktisere en ganske liberal tolkning av forskjellige varianter av bokmål uten hensyn til om de er konsekvent gjennomført (se Kulbrandstad, 2005, s. 69). Naturligvis vil en del feilretting være gjenstand for subjektivitet, men jeg vil følge to prinsipper og en veiledning: *det pragmatiske probabilitetsprinsippet* og *det minimale modifikasjonsprinsippet* for Norsk andrespråkskorpus ASK, som beskrevet i Tenfjord, Hagen og Johansen (2009, s. 60). Kort forklart betyr de to prinsippene at man velger det man tror *eleven har ment* og det som gir *minst endring* i teksten. Videre brukte jeg veiledningen for feilretting hos Polio og Shea (2014), som beskriver hvordan feil skal telles og korrigeres, med eksempler. Jeg beskriver både prinsippene og veiledningen i detalj i vedlegg 2. Prinsippene fra Tenfjord et al. (2009) vil noen ganger kollidere, men det skjedde så sjelden i min undersøkelse (under ti ganger totalt) at jeg tror det spiller en marginal rolle for resultatet. Som i all annen forskning har jeg gjort noen valg, men det er viktig å gjøre

disse så eksplisitt som mulig, med tanke på eventuell kritikk og gjentakelse senere (Polio, 1997, s. 129). Ved hjelp av et korrektkorpus av mine tekster ser jeg hvordan en målspråksbruker ville skrevet tekstene, iallfall på tekstlig mikronivå. For å styrke reliabiliteten skulle en annen målspråksbruker gjort det samme, slik at de kunne sammenlignes, men det har jeg ikke hatt mulighet til, det vil koste for mye, både i tid og penger.

3.2.7 Forskningsetiske betraktninger

Undersøkelsen ble gjort ved en anonymisert skole med nærmere 200 elever. Elevene er anonymisert. Ingen av elevene er den eneste fra sitt land på skolen, og jeg oppgir ikke hvilket førstespråk de har. Jeg kobler heller ikke opprinnelsesland og kjønn på deltakerne, og oppgir ikke hvem som har fått hvilke resultater. Ut fra denne besvarelsen skal det derfor ikke være mulig å spore deltakerne. Jeg har en koblingsnøkkel, men bare på papir, og den blir destruert når undersøkelsen er avsluttet. Sensitive opplysninger blir ikke samlet inn. Prosjektet er meldt til NSD og godkjent. Rektor og deltakerne ble orientert muntlig og skriftlig og undertegnet skjema for deltakelse (se vedlegg 3 og 4).

Jeg brukte et kvasieksperimentelt én-prøvedesign og gjorde noen forbedringer av designet for å få de usystematiske variasjonene så små som mulig. Elevene ble delt i to balanserte grupper med utgangspunkt i den uavhengige variabelen skrivemodus. Bakgrunnsvariablene antas ikke å ha betydning for undersøkelsen. 47 elever deltok, men av ulike årsaker ble det igjen 19 elever i hver gruppe. De håndskrevne tekstene ble digitalisert før jeg summerte to variabler. Deretter ble tekstene korrigert og feil ble telt etter noen prinsipper for korrigering og feiltelling. Jeg har anonymisert skolen og elevene og samler ikke inn sensitive opplysninger. Koblingsnøkkelen blir destruert når undersøkelsen er avsluttet. Prosjektet er meldt til NSD og godkjent, og elever og rektor er orientert og har undertegnet skjema for deltakelse. Jeg skal nå gjøre rede for hvordan kvalitet i skriving kan måles, før jeg gjør rede for leksikalske og syntaktiske variabler som jeg har brukt i mitt prosjekt.

3.3 Variabler for kvalitet i skriving

Ett av poengene med dette masterprosjektet var å finne forskjeller i skrivekvaliteten mellom de to modusene penn og papir og PC hos voksne andrespråksskrivere. Det gjorde jeg ved å anvende relevante syntaktiske og leksikalske variabler på elevtekster og sammenligne dem statistisk for de to modusene. De syntaktiske variablene var *gjennomsnittlig lengde av t-enhet*

(GTEL) og *lesbarhetsindeks (LIKS)* for å måle kompleksitet. For nøyaktighet brukte jeg *antall feil per t-enhet (FTE). Tekstlengde (TL)* viser om en elev har god flyt i sin skriving. De leksikalske variablene var *gjennomsnittlig ordlengde (GOL),* fordi jeg antok at dette viser størrelsen på ordforrådet. *Malvern og Richards' D (vocd D*) og *measure of linguistic textual density (MTLD)* er nyere variabler som har vist seg uavhengige av tekstlengde. Også *ordvariasjonsforhold (OVF)* måler leksikalsk variasjon. *Modifisert TTR (MTTR)* korrelerer godt med holistiske vurderinger og skolenivå. *Leksikalsk tetthet (LT)* kan vise andre egenskaper ved ordforrådet, den kan skille mellom modusene og vise «overføring» fra mer uformell digital skriving. *Lesbarhetsindeks (LIKS)* er enkel å utføre og ble med som validering av syntaktisk kompleksitet og leksikalsk variasjon. I det følgende vil jeg gjøre rede for hvordan og hvorfor disse målene er valgt som avhengige variabler. Hensikten er ikke å gjøre rede for funnene i forskningen, men å vise at disse variablene kan brukes i min oppgave.

For å kunne sammenligne tekster på en mest mulig nøytral og objektiv måte, må tekstene vurderes mest mulig objektivt og nøytralt. Holistiske vurderinger gjort av sensorer og lærere og score på språkprøver er i stor grad basert på skjønn, og intersensor-reliabiliteten kan være variabel. Slike vurderinger er langt vanskeligere i andrespråksskriving enn i skriving på morsmål, da tekstens gode sider ofte kan bli borte i overflatiske trekk, sier Hamp-Lyons og Kroll (1996). Sensorer klarer antagelig ikke å skille nøyaktighet fra andre globale mål som lengde og innhold når de bruker holistiske mål, sier Polio og Shea (2014, s. 22). Samtidig er det viktig å legge merke til at sensorene på Norskprøven prioriterer svar på oppgaven og sammenheng i teksten foran ordtilfang og grammatikk (Moe, 2002, s. 205). Likevel ønsket jeg å se om det faktisk var en forskjell mellom de to skrivemodusene i min oppgave, i størst mulig grad uavhengig av vurderinger og subjektivt skjønn. Da var det naturlig å tenke mer i analytisk og kvantitativ retning. Analytisk vurdering ser på flere aspekter i teksten og kan gi mer detaljert informasjon. Derfor foretrekkes analytisk vurdering ofte av skrivespesialister, sier Weigle (2002, s. 114). Alex Housen, Folkert Kuiken og Ineke Vedder (2012, s. 8) hevder at objektive, kvantitative måleenheter er best i studier av andrespråk. De mener at valide, reliable og praktiske mål for ferdigheter i andrespråket er mål for *kompleksitet, nøyaktighet og flyt*. Også Richard Hudson (2009) argumenterer for objektive mål, særlig fordi de er sensitive for *fravær* av språklige konstruksjoner. Likevel har denne måten å vurdere tekster på fått kritikk, blant annet fordi «kommunikative» elementer i språkevnen ikke vurderes tilstrekkelig (Plakans, 2014, s. 1391).

Jeg valgte i størst mulig grad mål som korrelerte med holistiske vurderinger og skolenivå (Wolfe-Quintero et al., 1998, s. 7f). Det felles europeiske rammeverket for språk deler lingvistisk kompetanse i tre deler: leksikalsk, fonologisk og syntaktisk kompetanse (Utdanningsdirektoratet, 2011, s. 15). Andre steder i rammeverket kan en finne flere typer kompetanse. I en undersøkelse av skriftlige besvarelser er det likevel naturlig å se på syntaktiske og leksikalske variabler. Det finnes også andre trekk i skriftspråket som kan være aktuelle, for eksempel kohesjon, stil/register, makrostruktur og retorikk, men det ble for krevende for mitt formål å se på alle disse. Blant disse er det bare kohesjon som er like framtredende som grammatikk og leksikon i vurderingskriteriene til Norskprøven (Vox, 2012).

3.3.1 Syntaktiske variabler

Først vil jeg si litt om terminologien jeg vil bruke. Ordet «setning» brukes på ulike måter i forskjellige sammenhenger, og det engelske «sentence» er ikke nødvendigvis det samme som det norske «setning», som i seg selv kan være tvetydig. Når jeg i det følgende bruker «setning», mener jeg en grafisk setning, markert med stort skilletegn og stor bokstav. Det er også hensiktsmessig å definere en kortere produksjonsenhet, *t-enhet*, som jeg skal forklare mer om under. Denne består av en eller flere klaususer (engelsk: clause), som Kellogg W. Hunt (1965) og Polio (1997) definerer som «en struktur med et subjekt og et finitt verb».

Gjennomsnittlig lengde av t-enhet

I dette avsnittet vil jeg drøfte kompleksitet som mål på språkutvikling, og i den forbindelse også modenhet, fordi det var utgangspunktet for utviklingen av t-enheten. Hunt (1965) ville finne en sammenhengende og systematisk måleenhet for grammatiske strukturer som også skulle fange opp detaljer. Han studerte morsmålet hos barn på tre nivåer i skolen, og kritiserte den etablerte bruken av setningen som en enhet. Det tradisjonelle var å se på den grafiske setningen, markert med store skilletegn og stor bokstav. Dette ble ofte feil for å studere syntaktiske enheter, mente Hunt, fordi barn ofte har en tendens til å skrive flere sideordnede klaususer etter hverandre uten å avslutte grafisk. Dette siste framsto som like komplisert eller utviklet som en setning med flere fraser og underordning når man brukte setningen som en enhet. Hunts løsning var å definere en «minimal terminable unit» (*t-unit*, *t-enhet* på norsk) «…which includes one main clause plus all the subordinate clauses attached to or embedded within it…» (Hunt, 1965, s. 157). Disse *kan* avsluttes med punktum og stor bokstav, derfor betegnelsen «terminable» (Hunt, 1965, s. 37). Ofte vil t-enhet utgjøre en grafisk setning, men

mange ganger vil en setning inneholde flere t-enheter, og noen ganger vil det være store skilletegn i en t-enhet.

En fordel med t-enhet er at underordning vil beholdes. Da kan man bruke underordning som et mål på modenhet eller språkutvikling. Hvis det finnes konjunksjoner som koordinerer hovedsetninger, vil disse markere grensen mellom to t-enheter. Setningsemner vil ikke telles med som t-enheter ut fra definisjonen over, men kan ofte henges på den foregående. På norsk er det vanlig å gjøre unntak fra setningsdefinisjonen for konstruksjoner med verbet i imperativ (Kulbrandstad, 2005, s. 181), så det har jeg telt som en t-enhet, men konjunktiv har jeg holdt utenom. Bare i tre tilfeller inneholdt mine tekster konjunktiv.

Senere inkluderte Hunt både ikke-klausale strukturer og setningsemner (Foster, Tonkyn & Wigglesworth, 2000, s. 360), men jeg har holdt meg til den opprinnelige definisjonen, fordi den er enklere å telle og viser bedre kontroll med målspråksgrammatikken i skriftlig norsk. Jeg har altså definert t-enheten som en hovedklausus pluss alle underordnete klaususer. Hunt så på den gjennomsnittlige lengden (målt i antall løpeord) av disse t-enhetene for å avdekke modent språk. Han fant at dette korrelerte bedre med klassetrinn enn gjennomsnittlig klaususlengde, underordningsforhold og setningslengde (Hunt, 1965, s. 39).

Senere prøvde Diane Larsen-Freeman å bruke gjennomsnittlig t-enhetslengde i tekstene til andrespråksinnlærere, men fant at de gjør for mange leksikalske, morfologiske og syntaktiske feil, og at man heller må se på feilfrie t-enheter fordi det også viser kontroll med syntaks på lik linje med lengre t-enhet (Larsen-Freeman, 1983, s. 288f). Gjennomsnittlig feilfri t-enhetslengde kunne måle kompleksitet, nøyaktighet og flyt (Larsen-Freeman, 2009, s. 580). Også Rod Ellis og Gary Barkhuizen (2005, s. 155) sier at t-enhet måler klausal underordning og fint kan brukes i analyse av kompleksitet i skriftlige andrespråkstekster.

15 år senere gikk Wolfe-Quintero et al. (1998) i en stor rapport gjennom over hundre måleenheter brukt i 39 studier i andrespråksskriving fram til da for å finne de mest lineære målene og de som korrelerte best med program og nivå i skolen og holistiske vurderinger (se også Hudson, 2009, s. 352). Med *lineære mål* menes et mål som utvikler seg jevnt og parallelt med språkutviklingen fram til full målspråkstilegnelse (Wolfe-Quintero et al., 1998, s. 2f). De fant at alle måleenhetene kunne klassifiseres innenfor flyt, nøyaktighet og kompleksitet, både grammatisk (syntaktisk) og leksikalsk (Wolfe-Quintero et al., 1998, s. 4). Senere fikk disse konstruktene akronymet CAF, for *complexity, accuracy and fluency* (Norris & Ortega, 2009).

Wolfe-Quintero et al. konkluderte med at de beste målene for flyt var t-enhetslengde, gjennomsnittlig lengde for feilfrie t-enheter og klaususlengde. For nøyaktighet var antall feilfrie t-enheter, andel feilfrie t-enheter og antall feil per t-enhet best, mens for syntaktisk kompleksitet var de beste måleenhetene antall klaususer per t-enhet, antall avhengige klaususer per totalt antall klaususer og antall avhengige klaususer per t-enhet. Begge de to siste måler klaususunderordning. I motsetning til Wolfe-Quintero et al. hevder Ortega (2003, s. 516) og Pekka Lintunen og Mari Mäkilä (2014, s. 387) at det er tradisjon for at lengdemål beskriver syntaktisk kompleksitet. Også Hunt (1965, s. 154) sier at *mer* er et tegn på kompleksitet.

I det hele tatt er konstruktene kompleksitet, nøyaktighet og flyt temmelig kompliserte og sammensatte, og de påvirker hverandre. Mye forskning har vist at de ikke samvarierer, og de utvikler seg i ulikt tempo. Skriving utvikler seg trinnvis, sier Christine Casanave (1994, s. 183). En diskusjon går mellom den såkalte «trade-off hypothesis», der innlæreren kan ha fokus på ett område, men at dette kan påvirke et annet område i negativ retning, og «cognition hypothesis» der økt kompleksitet i oppgaven fører til økt kompleksitet og nøyaktighet samtidig (Kuiken & Vedder, 2012, s. 147f; Skehan & Foster, 2012). Det er særlig kompleksitetskonstruktet som har vært mye omdiskutert, og de fleste undersøkelser har vært basert på t-enhet eller klaususunderordning (Biber, Gray, & Poonpon, 2011, s. 7). John M. Norris og Lourdes Ortega (2009) anbefaler at man bruker flere variabler for kompleksitet, de ser sannsynligvis på forskjellige aspekter av språket, hevder de. Derfor har jeg brukt tekstlengde (TL), gjennomsnittlig lengde av t-enhet (GTEL) og lesbarhetsindeks (LIKS) for å måle syntaktisk kompleksitet. Lesbarhetsindeks sier jeg mer om senere i kapittelet.

Som Norris og Ortega (2009) hevder Larsen-Freeman (1997; 2009) at kompleksitet er komplekst og argumenterer for at andrespråkstilegnelse og mål for kvalitet må forstås i lys av kompleksitetsteori. Begge er komplekse, dynamiske og ikke-lineære systemer, hvor faktorer påvirker hverandre. Faktorene påvirkes også av faktorer utenfor systemet og av startpunktet, dessuten forandres systemet hver gang det brukes, og små deler likner på større deler, sier hun. I tillegg er språk sosialt situert. Slik mener hun at den typiske innlæreren ikke finnes, at karakteristikker for grupper ikke kan brukes på enkeltpersoner og at vi må slutte å snakke om et målspråk. Likevel har forskningen etablert noen fellesnevnere. På 1990-tallet fikk de tre konstruktene (CAF) sine definisjoner, som holder ennå i dag: *Kompleksitet* forstås som «evnen til å bruke et bredt og variert utvalg av sofistikerte strukturer og vokabular i andrespråket», *nøyaktighet* som «evnen til å produsere målspråkslikt og feilfritt språk», og *flyt* som «evnen

til å produsere andrespråket med innfødt-lik hastighet, pauser, nøling og reformulering» (Housen et al., 2012, s. 2, egen oversettelse). Vi ser at de to siste sammenligner med målspråksbrukere, noe som i seg selv er problematisk (Polio, 1997).

En rekke andre måleenheter for modenhet i språk har blitt brukt (se Foster, Tonkyn og Wigglesworth (2000) for et utvalg), men t-enheten har blitt stående som den mest brukte syntaktiske måleenheten i analyse av språklige data (ibid.). Den er likevel ikke uten problemer. Foster et al. (2000) hevder at andrespråksbrukere planlegger produksjonen i kortere sekvenser, og innfører en kortere enhet, «Analysis of speech»-enhet. Den består av en uavhengig klausus eller subklausus-enhet, sammen med underordnede klaususer, slik at feil start, unyttige repetisjoner og selvkorrigering telles som separate enheter (Foster et al., 2000, s. 365). Dette er nok mest nyttig i tale, men hvis man leser andrespråkstekster, kan man finne dette også i skriftlige arbeider. Sandra Ishikawa (1995) er også inne på dette, men vil bruke klausus som en produksjonsenhet for andrespråkselever med lave ferdigheter og mange feil fordi den er mindre enn t-enheten og viser økning i ferdigheter på en bedre måte.

På den andre siden står Kathleen Bardovi-Harlig (1992), som mener at oppdeling av en setning i t-enheter forstyrrer en korrekt og retorisk viktig koordinasjon og underordning. Hun vil heller bruke setningen, som er en psykologisk og pedagogisk mer riktig produksjonsenhet for voksne andrespråksinnlærere, hevder hun. Douglas Biber et al. (2011) støtter hennes kritikk av bruken av t-enheten, men det er særlig fordi klaususunderordning ikke er typisk for avansert akademisk skriving, hevder de. Kritikken fra Biber et al. (2011) treffer likevel ikke helt, for slett ikke alle forskere de nevner i sin artikkel har undersøkt avansert akademisk skriving. Mine tekster hadde heller ikke dette som mål. Det er også, som nevnt, slik at kompleksitet ikke utvikler seg lineært, men kan påvirkes og svekkes av utvikling i andre konstrukter. Michael A. K. Halliday (1989, s. 62f) går så langt som til å si at skriftspråket er *leksikalsk* komplekst, men mindre grammatikalsk komplekst enn talespråket. Også innenfor førstespråksforskning er det kritikk av bruken av t-enheten (se for eksempel Grabe og Kaplan, 1996; Hudson, 2009), men det viktigste i min sammenheng er *hvordan* den skal brukes, ikke *om* den skal brukes.

Antall feil per t-enhet

Når det gjelder feil, så vi at Wolfe-Quintero et al. (1998) fant at antall feilfrie t-enheter og antall feil per t-enhet korrelerte best med holistiske vurderinger og skolenivå. Jeg antok at det var vanskelig å finne mange feilfrie t-enheter i mine tekster, så da ble valget enkelt, jeg brukte

antall feil per t-enhet. Kuiken og Vedder (2012, s. 146) kommer til samme konklusjon for begynnere og litt erfarne skrivere. En ulempe med å bare telle feil er at alvorlighetsgrad og type feil ikke kommer fram, sier de. Polio og Shea (2014) vurderer validitet og reliabilitet for variabler brukt i 35 studier mellom år 2000 og 2011, og anbefaler heller ikke feilfrie t-enheter for begynnere. En liten feil fører til at hele t-enheten regnes som ugrammatisk. De finner ikke at ett mål for nøyaktighet er mer reliabelt eller valid enn andre, men tilrår at man er klar over styrker og svakheter ved de enkelte variablene. Løsningen er likevel ikke å bruke mer enn én variabel, hevder de, fordi variablene er mer redundante her enn i kompleksitets-konstruktet, alle måler antall feil på en eller annen måte (Polio & Shea, 2014, s. 22). I vurderingsskjemaet for sensorene på Norskprøven (Vox, 2016g), er *Rettskriving og tegnsetting* et eget kriterium, og jeg har snakket med flere sensorer som sier de legger vekt på feil. Vi vet også at feil påvirker inntrykket av teksten vi leser, feil er saliente. Jeg har brukt variabelen *antall feil per t-enhet (FTE).* Denne variabelen tar opp alle typer feil, ikke bare syntaktiske og leksikalske feil. I vedlegg 2 gir jeg en oversikt over feiltyper og alvorlighetsgrad i mitt prosjekt.

En skal være forsiktig med å overføre all engelskspråklig språkforskning ukritisk til andre språk (Matsuda, 2003, s. 28; Ortega, 2009, s. 232f). For eksempel er det velkjent at norsk har relativt mindre stilistisk avstand mellom skrift og tale enn mange andre språk, og vi blir ofte oppfordret til å skrive enkelt (se for eksempel Difi, 2016; Søyland & Fretland, 2015). Også morfologiske, syntaktiske og leksikalske forhold skiller seg fra engelsk. Mer komplekst trenger heller ikke nødvendigvis å være bedre, det hender at syntaksen er enkel på grunn av bevisste pragmatiske og stilistiske valg (Pallotti, 2009, s. 598). Dessuten kan for eksempel økt kompleksitet gi mindre nøyaktighet fordi disse konstruktene kan være avhengige av hverandre (Plakans, 2014, s. 1393), noe som rent intuitivt virker sannsynlig. Det er gjort relativt lite kvantitativ forskning i Skandinavia på skrivekvalitet, men det finnes litt, og mest i Sverige. I Norge er det gjort noe skriveforskning, men det meste er kvalitativt, sier Sigmund Ongstad (2002). Etter Ongstads artikkel har det økt litt, og Skrivesenteret ved NTNU er etablert.

Tekstlengde

De svenske prosjektene *Talsyntax* og *Skrivsyntax* på 1970-tallet etablerte et solid grunnlag for kvantitativ forskning på elevtekster skrevet på morsmålet. En del av *Skrivsyntax* var *Gymnasistsvenska*, hvor Tor G. Hultman og Margareta Westman (1977) fant en klar korrelasjon mellom tekstlengde og karakter, det var bare den øverste karakteren som ikke kunne bestemmes av en datamaskin, sier de (Hultman & Westman, 1977, s. 54). Også Kent

Larsson (1984, s. 192) fant i sitt doktorgradsprosjekt en sterk korrelasjon (0,82) mellom antall ord og karakter for elever mellom 13 og 15 år. Dessverre oppgir han ikke noe mer om denne korrelasjonskoeffisienten. Det er elevens evne og vilje til å uttrykke seg og bruke språket som viser seg i tekstlengden. Produktiviteten er «…ett symptom på en språkbehärskning som yttrar seg i att man kan använda språket för att lösa den uppgift man har fått» (Hultman & Westman, 1977, s. 55). I sin delstudie innenfor prosjektet «Kvalitetssikring av læringsutbyttet i norsk skriftlig» bruker Wenche Vagle (2005) også tekstlengde som en variabel og finner en korrelasjon med karakter (Pearsons *r* for de forskjellige eksamensårene varierer mellom 0,42 og 0,52).

Innenfor andrespråk rapporterer flere lignende korrelasjoner i oversikter over tidligere forskning (Jarvis, Grant, Bikowski & Ferris, 2003; Larsen-Freeman, 1978; Ortega, 2009). Lengre tekster er en generell effekt av tekstbehandling, sier Pennington (2003, s. 289) i en oppsummering av flere studier på andrespråkselever. Jeg antok at tekstlengde ville skille tekster skrevet med penn fra tekster skrevet på PC, iallfall for elever som hadde god nok skrivehastighet på PC. I sitt upubliserte doktorgradsarbeid finner Bård Uri Jensen at 50 av hans 60 elever i VG1 skriver lengre på PC enn for hånd[15], men det er sannsynlig at hans norskspråklige elever hadde bedre PC-ferdigheter enn mine. I mitt prosjekt er det også viktig å ta hensyn til at oppgavene på Norskprøven anbefaler en ramme for tekstlengde, for eksempel 80-200 ord, slik som det er i oppgave 2 for nivå A2/B1 og i min prøve. Dermed kan det hende at elevene stopper når de har skrevet «nok». Jeg har brukt variabelen *tekstlengde (TL).*

3.3.2 Leksikalske variabler

Det er ingen grunn til å tvile på at ordforrådets størrelse er viktig for skrivekvaliteten i et andrespråk. Batia Laufer og Paul Nation (1995, s. 307f) refererer flere undersøkelser som viser at en rekke leksikalske mål korrelerte positivt med holistiske vurderinger av skrivekvalitet. James Milton (2010, s. 225) har sett på utviklingen av ordforrådet koblet til nivåene i Rammeverket, og finner at 60 til 70 % av variansen mellom nivåene kan forklares med forskjeller i ordforrådet (hvis man fjerner ungarsk). Endelig sier Rønnaug Katharina Totland og Hanne Lauvik i prosjektet *Norsk profil* at «ordforrådet er det som spiller størst rolle for både lese-, lytte-, snakke- og skriveferdigheter» (Totland & Lauvik, 2012, s. 189). Det er flere

[15] Personlig kommunikasjon, e-post juni 2016.

måter å måle leksikon på, jeg skal nå drøfte noen av dem. *Feil* i ordforråd dekkes til en viss grad gjennom variabelen *feil per t-enhet*, som jeg viste tidligere.

Gjennomsnittlig ordlengde

Ordlengde er en variabel som kan predikere skrivekvalitet, sofistikerte ord er ofte lengre enn enkle ord. Ordlengde kan også si noe om morfologien eller ords opphav, bøyde ord eller fremmedord er ofte lengre (Golden, 2014, s. 186). Vagle (2005) bruker dette og sier at det er en sammenheng mellom ords lengde, frekvens og semantiske «tyngde» (Vagle, 2005, s. 305). Dette kalles «Zipfs lov» (som Gustav Herdan (1964) sier er verken Zipfs eller en lovmessighet). Hultman og Westman (1977) i prosjektet *Gymnasistsvenska* og Eva Östlund-Stjärnegårdh (2002) i sin doktoravhandling finner en svak, og litt utydelig sammenheng mellom gjennomsnittlig ordlengde per tekst og karakter i svenske gymnaser. Også Vagle (2005, s. 350) finner at korrelasjonen er svakere enn for tekstlengde, men fortsatt relativt tydelig (Pearsons *r* fra 0,34 til 0,40 for de forskjellige årene hun har undersøkt).

Innenfor andrespråksfeltet finner Scott Jarvis et al. (2003) at svake og sterke tekster skiller seg klart fra hverandre i ordlengde. I Skandinavia kjenner jeg bare til ett prosjekt som bruker flere leksikalske variabler på andrespråkstekster. Det er Kokkinakis og Magnusson (2011) i en delstudie i prosjektet *Språk och språkbruk hos ungdomar i flerspråkiga storstadsmiljöer (SUF)* ved Universitetet i Göteborg. De bruker fire kvantitative variabler: *leksikalsk tetthet* (som jeg forklarer senere), *nominalforhold* (andelen nominaler i teksten), *ordvariasjonsindeks* (en logaritmisk transformasjon av TTR - se under) og *ordlengde*[16]. De fant at ordlengde korrelerte best med karakter og de andre variablene, og kunne dermed predikere de andre variablene. Tre av variablene hos Kokkinakis og Magnusson (2011) er identiske med eller ligner på mine variabler, nominalforhold hadde jeg ikke med. Jeg har brukt variabelen *gjennomsnittlig ordlengde (GOL).*

Når det gjelder leksikalsk variasjon, ser det enkelt ut å dele antall forskjellige ordtyper (engelsk: types) på totalt antall ord (tokens) i en tekst. Slik kan man få et forholdstall (ratio). Problemet er imidlertid at forholdstallet faller når tekstene blir lengre fordi man bruker ord på nytt, telleren i brøken øker mer enn nevneren. Kurven får form som en hyperbel og nærmer seg null med økende antall ord, men likevel slik at tekster med stor leksikalsk variasjon ligger

[16] Navnene på målene er mine oversettelser fra engelsk.

over tekster med liten variasjon. Fordi TTR (type-token ratio) ikke er konstant med økende tekstlengde kan den ikke brukes til å sammenligne tekster med forskjellig lengde (McKee, Malvern, & Richards, 2000). Dette er et problem for reliabiliteten til variabelen (Broeder, Extra & van Hout, 1993, s. 148). Aneta Dewaele og Jean-Marc Pavlenko (2003, s. 129) viser tre måter å løse problemet på: 1) å bruke utdrag av lik lengde (vanligvis 1000 ord), 2) å bestemme andelen lavfrekvente ord i et utdrag eller 3) å bruke en formel som er mest mulig tilpasset TTR-kurven. De to første var uaktuelle i min oppgave fordi jeg ikke hadde lange nok tekster og fordi elevene i liten grad brukte lavfrekvente ord. Da sto jeg igjen med den siste, å bruke en formel som er mest mulig tilpasset TTR-kurven. Jeg skal nå se på noen slike formler.

Malvern og Richards' D (vocd D)

Helt siden 1930-tallet har det vært gjort forsøk på å finne en formel for leksikalsk variasjon som ikke påvirkes av tekstlengden (McCarthy & Jarvis, 2007). Jarvis (2002) diskuterer en del av de formlene som er vanlige: *Herdans C, Guirauds R, Zipfs Z, Malvern og Richards' D* og *Dugasts Uber*. Han finner at de to sistnevnte passer best til tekster av forskjellig lengde (Jarvis, 2002, s. 81). Pilar Durán, David Malvern, Brian Richards og Ngoni Chipere (2004) hevder at Malvern og Richards' D er best egnet til å beskrive leksikalsk variasjon (D betyr *diversity)*. Denne variabelen betegnes også *vocd D*[17] og har blitt noe av en «industristandard» de senere årene, sier Philip McCarthy og Scott Jarvis (2007), særlig ved at det ble brukt i CHILDES-prosjektet og lagt åpent på internett (MacWhinney, 2003). Variabelen vocd D finnes ved at 35 ord tas tilfeldig fra teksten. Dette gjøres 100 ganger, et gjennomsnitt regnes ut og man får et punkt i et koordinatsystem. Deretter tar man 36 ord tilfeldig ut av teksten, og det samme gjentas. Slik fortsetter det til man når 50 ord og sitter med 16 punkter som danner en kurve. Denne tilpasses teoretiske kurver som er laget tidligere fra mange forsøk. Til slutt får man en kurve som passer best mulig til den faktiske TTR (Durán et al., 2004). Målet gir ikke mening i seg selv, men må sammenlignes med andre.

Som nevnt, fant Jarvis (2002) at vocd D passet godt til tekster av forskjellig lengde. Senere kritiserer McCarthy og Jarvis (2007) vocd D for at den ikke er uavhengig av tekstlengde, den gir litt forskjellig resultat hver gang, men særlig er det samplingen som er problemet, hevder de. Ved at vocd sampler fra hele teksten hver gang den tar et utvalg, kan egentlig hvert punkt

[17] Vocd er et dataprogram skrevet av Gerard McKee ved University of Reading (McKee et al., 2000).

på kurven representere hele teksten (McCarthy & Jarvis, 2007, s. 468). De mener likevel at målet kan brukes på tekster mellom 100 og 400 ord, så jeg fjernet tekster utenfor denne rammen. Jeg har brukt variabelen vocd D.

Measure of textual lexical diversity

Senere lanserte McCarthy *Measure of textual lexical diversity (MTLD)*, og denne variabelen er gjenstand for en grundig gjennomgang hos McCarthy og Jarvis (2010). De sammenligner med en rekke variabler som er brukt i de senere årene (vocd D, TTR, Maas, Yules K og HD-D) og finner at MTLD er den eneste variabelen som ikke påvirkes av tekstlengde. De validerer også MTLD grundig. Algoritmen for MTLD er slik at den finner det gjennomsnittlige antall ord som er skrevet før TTR stabiliseres. Den analyserer teksten sekvensielt, med ett og ett nytt ord, og regner ut samlet TTR for hvert nye ord. Når TTR når 0,720, telles antall ord og faktoren settes til 1 igjen før algoritmen begynner på nytt til teksten er analysert og man får et gjennomsnittlig antall ord. Det hele gjentas bakfra i teksten og det regnes ut en middelverdi av de to. Dette gir MTLD. Tallet 0,720 er valgt empirisk ved å undersøke en rekke tekster og finne hvor kurven for TTR når et «metningspunkt», hvor verken repetisjon av ord eller en mengde nye ord påvirker kurven nevneverdig (McCarthy & Jarvis, 2010, s. 386). MTLD kan ikke brukes på tekster under 100 ord. McCarthy og Jarvis (2010) viser også at vocd D, MTLD og Maas fanger opp forskjellige deler av lingvistisk variasjon. Nettsiden *textinspector.com* regner ut både vocd D og MTLD, men jeg erstattet alle ikke-engelske tegn med engelske. MTLD er en måleenhet og gir ikke mening i seg selv. Jarvis (2013) viser at teksten som helhet ikke vurderes av noen av disse målene, så hvis tre avsnitt er like, vil det ikke slå ut, verken for vocd D eller MTLD. Det ble da min oppgave å finne ut av. Jeg har brukt variabelen MTLD.

Ordvariasjonsforhold

Også i Sverige ble det tidlig konstruert alternativer til TTR. Hultman og Westman (1977) brukte «ordvariationsindex» (OVIX), der tekster med mange ulike ord i forhold til tekstlengde får en høyere OVIX-verdi enn tekster med få ulike ord (Hultman & Westman, 1977, s. 56). Problemet med OVIX (for meg) er at det ikke kan brukes på tekster under 200 ord (Nyström, 2000, s. 177). Nettstedet lix.se regner ut OVIX og et annet mål, «ordvariationsratio» (OVR). Formelen for OVR lages ved å ta logaritmen for ordtyper (types) og dele på logaritmen for antall ord i teksten (tokens), da får man en kurve som heller stiger sakte og den skal passe bedre for korte tekster (Nyman, 2014; Seimyr, udatert). OVR er egentlig Herdans C, som vi ser hvis vi går inn på formelen (Egghe, 2007, s. 702; Jarvis, 2002, s. 71). Selv om Herdans C

ikke passer til alle tekstlengder, følger den kurven for faktisk TTR ganske godt for tekster i området fra omtrent 85 ord til omtrent 170 ord (Jarvis, 2002, s. 72). Gerard McKee et al. (2000) viser hvorfor TTR likevel ikke kan brukes, den er ikke reliabel, men gir forskjellig tall ved ulike tekstlengder. De fleste av mine tekster lå i området 85 til 170 ord. McCarthy og Jarvis (2010) hevder at variabler som bruker logkorreksjon (som OVR) bare påvirkes 1,5% av tekstlengde. Kokkinakis og Magnusson (2011, s. 112) fant i en forstudie til sitt prosjekt innen SUF ikke at OVR korrelerte signifikant med andre variabler, så de brukte en mer komplisert variant. De oppgir ikke hvilke variabler de ikke fant korrelasjon med, så jeg tar ikke hensyn til dette. Det var interessant å ha med et mål som er brukt i skandinavisk forskning i tillegg til de engelskspråklige, men jeg har fornorsket selve uttrykket litt og brukt variabelen *ordvariasjonsforhold* (OVF).

Wolfe-Quintero et al. (1998, s. 119) fant at de beste variablene for leksikalsk kompleksitet var antall sofistikerte ordtyper per totalt antall ordtyper og ordtyper per kvadratroten av to ganger totalt antall ord. Det førstnevnte er algoritmen Laufer (1994) utviklet til sin *leksikalske frekvensprofil*. Den måler hvor stor andel av ordene i teksten som er sofistikerte, i motsetning til et variasjonsmål, som bare måler hvor mye ordvariasjon en skriver har i teksten (Wolfe-Quintero et al., 1998). Ulempen med Laufers variabel er at en må ha tilgang på en liste over sofistikerte ord (Golden, 2014, s. 89). Monsen (2008) refererer betydelig kritikk av denne modellen fra flere forskere, og hun finner heller ikke at bestått besvarelser til Norskprøve 3 har større leksikalsk rikdom enn de ikke-bestått. For min egen del vil jeg legge til at jeg synes det virker uheldig med en dikotomisering av ordene, at de enten er inne eller ute på en liste. Det er også sånn at ord i mange tilfeller er domene- og sjangerspesifikke. Jeg mener også at jeg fanget opp sofistikasjon til en viss grad ved å bruke ordlengde i flere av de andre variablene.

Modifisert TTR

Den andre variabelen hos Wolfe-Quintero et al. (1998) er hentet fra Bradford Arthur (1979). Han bruker en modifisert variant av TTR, som nevnt over, og finner signifikant forbedring i leksikalsk variasjon hos elever med lave til middels språkferdigheter i løpet av et åtteukers kurs. Dette målet fanger opp både ordtyper og tekstlengde, slik at en elev med lengre tekst får et høyere tall. Det er også uavhengig av ordlister og dataprogrammer. Modellen er hentet fra John B. Carroll (1967). Jarvis (2013) viser at heller ikke denne modellen er uavhengig av

tekstlengde, men siden den korrelerte så godt hos Wolfe-Quintero et al. (1998), tok jeg den med. Jeg har kalt denne variabelen *MTTR*.

Leksikalsk tetthet

Forskningen på ordforråd skiller mellom *grammatiske ord (funksjonsord)* og *innholdsord (leksikalske ord)*. Det er innholdsordene som bærer det meste av meningen i teksten, uten dem forstår vi ikke mye av innholdet. Innholdsordene regnes vanligvis som ordklassene *substantiv, verb (unntatt hjelpeverb), adjektiv* og *adverb som er avledet av adjektiv* (Golden, 2014, s. 42). De grammatiske ordene består av de andre ordklassene, og viser ofte bare forholdet mellom innholdsordene. Andrespråkselever lærer ofte funksjonsordene senere enn innholdsordene, og noen ganger mangler en del funksjonsord i andrespråkstekster (Martens, 2009). Skillet mellom grammatiske ord og innholdsord er likevel ikke så uproblematisk som det kan synes, viser Golden (2014, s. 43). Jeg drøfter avgrensningen her i vedlegg 5. Leksikalsk tetthet regnes ut ved å ta antall eksemplarer (token) innholdsord i teksten og dividere med antall ordeksemplarer (token) totalt i teksten, multiplisert med 100 hvis man ønsker et prosenttall (Golden, 2014, s. 87). Færre grammatiske ord og/eller flere innholdsord gir høyere leksikalsk tetthet. Dette er avhengig av modus (tale/skrift), sjanger og kontekst (Johansson, 2008), men også planleggingstid og redigeringsmuligheter vil sannsynligvis virke inn på leksikalsk tetthet. Dermed kan det være forskjell i leksikalsk tetthet mellom PC og penn og papir, fordi redigeringsmulighetene er forskjellige i de to modusene.

Nå er det slik at andrespråkselever sannsynligvis kjenner færre synonymer og mange har ikke utviklet et mer akademisk språk. Dette kaller Jim Cummins (1981; 2008) «cognitive/academic language proficiency» (CALP) og sier at det tar minst fem år i gjennomsnitt å oppnå dette nivået. De fleste av mine informanter har lært norsk i kortere tid enn fem år, og det er sannsynlig at de vil bruke et mer grunnleggende vokabular (basic interpersonal communication skills – BICS). Dermed er det ikke sikkert at de som får bedre tid og andre redigeringsmuligheter når de skriver på PC vil få økt leksikalsk tetthet likevel.

Leksikalsk tetthet kan vise kvalitet i skriving, men Wolfe-Quintero et al. (1998, s. 105) finner at dette ikke har noe med andrespråksutvikling å gjøre, det støttes av Anne Marit Danbolt (2004, s. 4), som siterer Moira Linnarud. Cheryl A. Engber (1995) hevder at leksikalsk tetthet ikke har sammenheng med holistisk vurdering i form av karakterer hos andrespråkselever. Også Kokkinakis og Magnusson (2011) finner at leksikalsk tetthet ikke korrelerer med andre mål eller gir signifikante forskjeller ved forskjellige karakterer. Kenneth Hyltenstam (1988)

viser også at det er mulig å få høy tetthet med lite vokabular, eller ved repetisjoner (Linnarud, 1986). Likevel antok jeg at leksikalsk tetthet kunne være interessant å se på, for det er sannsynlig at leksikalsk tetthet henger sammen med blant annet modus, som jeg har undersøkt. Det er også velkjent at talespråk har lavere leksikalsk tetthet enn skriftspråk (Halliday, 1987; Ure, 1971), og skriving på PC kan i noen sammenhenger ligne på muntlig språk (Biber & Conrad, 2009; Collot & Belmore, 1996). Jeg trodde også at noen elever kunne «overføre» skrivevaner fra uformell skriving med digitale verktøy til en prøvesituasjon, iallfall de som skriver ofte og fort på slike verktøy. Slik uformell skriving (tekstmeldinger) ligner litt på muntlig språk, men har flere verb og pronomener, omtrent like mange substantiver, men betraktelig færre adverb og adjektiver enn andre typer digital kommunikasjon (Biber & Conrad, 2009, s. 207). Hvordan dette slår ut i leksikalsk tetthet er usikkert, men kan være interessant å undersøke. Jeg har brukt variabelen *leksikalsk tetthet (LT).*

En kombinert variabel

I Sverige er også *lesbarhetsindeks* mye brukt, utviklet av Carl-Hugo Björnsson (1968). Det måler både syntaktisk kompleksitet, leksikalsk kompleksitet, leksikalsk variasjon og leksikalsk sofistikasjon ved at gjennomsnittlig setningslengde og andel lange ord (over seks bokstaver) summeres. Det kan bare brukes som en pekepinn og har fått mye kritikk, blant annet fordi også uvanlige ord kan være korte, ord med et bøyningsmorfem framstår som vanskeligere enn uten, og at venstretunge setninger bedømmes som like vanskelige som setninger der subjekt og predikat er nær hverandre. Björnsson advarte selv mot dette (Lundberg & Reichenberg, 2009). I andrespråkssammenheng kommer også problemer med lange setninger, svak tegnsetting og feilstaving inn, slik at tekstene må korrigeres først, som jeg har vært inne på før. Som OVF er dette et skandinavisk mål, som kan være interessant å bruke på andrespråkstekster. Jeg har brukt *lesbarhetsindeks (LIKS).*

Oppsummering

Objektive, kvantitative variabler er anvendelige og beskrivende for skrivekvalitet i andrespråket. Av disse har jeg holdt meg til syntaktiske og leksikalske variabler. Mål for kompleksitet, nøyaktighet og flyt er valide, reliable og praktiske (Housen et al., 2012, s. 8). Jeg har benyttet gjennomsnittlig lengde av t-enhet (GTEL) for å måle kompleksitet. For nøyaktighet har jeg brukt feil per t-enhet (FTE), ifølge Polio og Shea (2014) er det nok med én variabel som måler feil. Jeg antok at tekstlengde (TL) viste om en elev hadde god flyt i sin skriving. Lange ord viser om eleven utnytter ordforrådet sitt forskjellig i to moduser, derfor

ble gjennomsnittlig ordlengde (GOL) mitt valg her. Malvern og Richards' D og MTLD er nyere variabler for leksikalsk variasjon som i liten grad er avhengige av tekstlengde. Ordvariasjonsforhold (OVF) måler leksikalsk variasjon, som kan øke når eleven får bedre tid eller redigeringsmuligheter. Arthur (1979) bruker en modifisert TTR (MTTR) som korrelerer godt med holistiske vurderinger og skolenivå. Den kunne også fange opp svakheter i flere av de andre variablene. Leksikalsk kompetanse og feil er også viktig i kommunikasjon og i Norskprøven (Monsen, 2008). Leksikalsk tetthet (LT) kunne vise modusforskjeller og «smitte» fra uformell digital skriving. Lesbarhetsindeks (LIKS) var enkel å utføre og ble med som validering av syntaktisk kompleksitet og leksikalsk variasjon. Jeg fulgte rådet fra Polio og Shea (2014) og Bulté og Housen (2014) om å bruke flere variabler i vurderingen av kvalitet.

I alt brukte jeg ti variabler, tre syntaktiske, seks leksikalske og én kombinert variabel (se tabell 1). Selv om det var betraktelig flere leksikalske enn syntaktiske variabler, ble de ikke satt opp mot hverandre i analysen. De leksikalske variablene beskrev også forskjellige kvaliteter ved ordtilfanget. Også i måling av ordforråd er det viktig å bruke flere variabler, sier Golden (2014). Fordelingen mellom syntaktiske og leksikalske variabler gjenspeiler påstandene hos Milton (2010) og flere om at ordforrådet er viktigst for å forklare andrespråksutvikling, som nevnt tidligere. Tabell 1 oppsummerer variablene jeg har drøftet i dette kapittelet.

Tabell 1. Syntaktiske og leksikalske variabler, forkortelser og formler/nettsteder

Full benevnelse	**Forkortelse**	**Formel/nettsted**
Gjennomsnittlig t-enhetslengde	GTEL	Antall løpeord i teksten/ antall t-enheter i teksten
Feil per t-enhet	FTE	Antall feil i teksten/ antall t-enheter i teksten
Tekstlengde	TL	Antall løpeord i teksten
Gjennomsnittlig ordlengde	GOL	Antall tegn i teksten/ antall løpeord i teksten
Malvern og Richards' D	vocd D	textinspector.com*
Measurement of textual lexical diversity	MTLD	textinspector.com*
Ordvariasjonsforhold	OVF	Log $type_n$/ log $token_n$
Modifisert TTR	MTTR	$Type_n$/ $\sqrt{2}$ $token_n$
Leksikalsk tetthet	LT	Innholdsord/ løpeord i teksten*100
Lesbarhetsindeks	LIKS	lix.se*

**Algoritmen for disse er forklart i teksten*

3.4 Statistiske metoder og valg

Til den statistiske behandlingen av data har jeg brukt IBM SPSS 23 (IBM, 2015). Jeg skal vise metoder og valg jeg har gjort i den forbindelse. Først viser jeg hvordan jeg har korrigert for mange tester, deretter hvilke testpremisser som er lagt til grunn, før jeg redegjør for *p*-verdier, effektstørrelser og statistisk styrke i testene jeg har brukt.

3.4.1 Korreksjon for mange tester

Jeg hadde altså ti variabler for kvalitet i skriving, variabler som sannsynligvis beskrev deler av de ulike konstruktene flyt, nøyaktighet og kompleksitet. Da hadde det vært nyttig å gjennomføre en prinsipalkomponentanalyse og deretter en multippel regresjonsanalyse for å finne ut hvilke variabler som bidrar mest til konstruktene. Dette er nok noe utenfor et masterprosjekt, så jeg valgte å kjøre uparete t-tester for alle de uavhengige variablene (Mann-Whitney U-test for ikke normalfordelte observasjoner). Problemet med å kjøre gjentatte tester som er avhengige av hverandre eller tilhører samme konstrukt er at sjansen for å finne «falske positive» øker, vi kan begå en type 1-feil. Hvis vi setter alfa-verdien til 0,05 har vi på grunn av naturlig variasjon 5 % sjanse for å finne noe som ikke er der. Denne øker for hver gang vi kjører en test, slik at med ti tester har risikoen for å finne en «falsk positiv» økt til over 40 %[18]. Det er selvfølgelig ikke akseptabelt.

Det finnes flere måter å korrigere dette på, blant annet Bonferroni-korreksjon. Den er ganske streng (Larson-Hall, 2010, s. 390), ved at den antar at sjansen for å gjøre en type 1-feil er over 40 % fra første test og deler dermed alfa-verdien (0,05) på antall tester (her 10) slik at den blir på 0,005. Dette øker sjansen for å gjøre en type 2-feil mye, det vil si at vi ikke finner noe signifikant som faktisk er der. Derimot vil en Holm-Bonferroni-korreksjon kontrollere den laveste p-verdien mot den strengeste alfa-verdien ved å gå sekvensielt fram (Holm, 1979). Med denne metoden må den laveste *p*-verdien være under 0,05/10 (i mitt tilfelle), den nest laveste under 0,05/9, den tredje under 0,05/8 og så videre. Når en signifikant verdi nås, kan en dermed si at resten også er signifikant. Med det øker sjansen for å finne noe signifikant, testen har altså mer styrke, samtidig som vi bedre unngår type 1-feil. Denne korreksjonen gjøres etter

[18] Nøyaktig 40,13 %, formelen er: $\alpha = 1 - (1 - \alpha)^c$, hvor c er antall tester (Hatch & Lazaraton, 1991, s. 263).

at de statistiske testene er tatt, og kan brukes på både parametriske og ikke-parametriske data (Holm, 1979).

Jeg har brukt et Excel-ark for utregningen av Holm-Bonferroni-korreksjonen og signifikansen (Gaetano, 2013). I en undersøkelse med mer avgrensete konstrukter og hypoteser og måleenheter som måler bare på ett konstrukt, kunne en hatt korreksjon for hvert konstrukt eller hypotese, for eksempel bare for kompleksitet, men i min undersøkelse henger konstruktene sammen og måleenhetene overlapper hverandre, så jeg har valgt å bruke Holm-Bonferroni-korreksjonen på alle testene under ett. Dette anbefales i statistisk litteratur (for eksempel Veazie, 2006). Signifikansen i testene blir oppgitt etter at denne korreksjonen er gjort.

3.4.2 Testpremisser

Det er fire forutsetninger for t-tester: den avhengige variabelen skal måles på intervallnivå, data skal være uavhengig, data skal være normalfordelt og gruppene skal ha homogen varians (Larson-Hall, 2010, s. 250). De to siste kan kontrolleres i SPSS. Normalfordeling er naturligvis ikke et premiss for Mann-Whitney U-test, men de andre tre er et premiss også for den.

Avhengig variabel på intervallnivå

En intervallskala har like lang avstand mellom alle tallene på skalaen, og null har ikke noen mening, som for eksempel en temperaturskala, målt i Celsiusgrader. Hvis null har en mening (altså at noe «mangler»), er det en forholdstallsskala (ratio scale), som farten på et prosjektil. En del av mine variabler er et forhold mellom to størrelser, og har et absolutt nullpunkt (for eksempel null feil), men flere forfattere sier at forholdstall ikke har noen betydning i anvendt lingvistikk og at de begge benevnes som intervallskalaer (Larson-Hall, 2010, s. 34; Lowie & Seton, 2013, s. 21). Mine avhengige variabler gir tall med lik distanse mellom seg. De er på intervallnivå.

Uavhengige data

Uavhengige data betyr i denne sammenhengen at resultatene vi får fra tekstene ikke er fra samme personer. I streng forstand er det en viss avhengighet mellom elevene i samme klasse, på samme skole, på samme bosted og så videre (Lie, Kjærnsli, Roe & Turmo, 2001, s. 82), men dette er en kvasi-eksperimentell undersøkelse, og de er ofte slik. Jeg har brukt to forskjellige grupper, hver person har skrevet enten på PC eller med penn, og vi kan gå ut fra at data er uavhengig.

Normalfordelte data

Parametriske tester har som premiss at fordelingen av verdiene tilnærmelsesvis følger en kjent matematisk fordelingsfunksjon som vi kjenner parametrene til. Hvis den ikke gjør det, kan ikke parametriske tester brukes. Jeg har undersøkt om data fra alle de avhengige variablene (i alt 20 distribusjoner) er normalfordelt ved visuell inspeksjon (histogram, boksplot, stem and leaf og Normal Q-Q-plot) og numerisk kontroll (Shapiro-Wilk test og skjevhet og kurtose) gjennom SPSS. Akseptable rammer for skjevhet og kurtose kan beregnes ved å dele verdien på standardfeilen (SE). Hvis denne Z-verdien er innenfor +/- 1,96, kan data anses å være normale. Dette kan brukes for utvalg under 100 (Fife-Schaw, 2016). Hvis verdiene over var innenfor anbefalte rammer, brukte jeg uparete t-tester for å finne forskjeller mellom gruppene, ellers brukte jeg uparete Mann-Whitney U-tester, som ikke har normal fordeling som premiss. Dette er gjort for alle de avhengige variablene, det vil si de syntaktiske og leksikalske variablene for kvalitet i skriving. Ikke alle mine data er normalfordelte, men da har jeg brukt en test som passer til ikke-normalfordelte data.

Homogen varians

Det fjerde premisset for normalfordelte data er at variansen mellom utvalgene er homogen. Jeg har brukt Levenes test i SPSS for å undersøke dette, selv om Jenifer Larson-Hall (2010, s. 88) sier at testen ikke bør brukes med små utvalgsstørrelser, uten at hun angir hva det betyr. SPSS gir verdier for både Students T-test og Welchs T-test, den siste er robust mot både homogen varians og ulike gruppestørrelser (Ruxton, 2006), så jeg har brukt den i tillegg. For ikke-parametriske data har jeg gjennomført en alternativ Levenes test, en parametrisk variansanalyse utført på rang-transformerte data (Nordstokke et al., 2011, s. 3). Også ikke-parametriske tester må ha homogen varians, sier David W. Nordstokke et al. (2011, s. 2) og Donald W. Zimmerman (2004, s. 1), selv om andre mener at dette ikke er nødvendig (Field, 2000, s. 49). Jeg har undersøkt om variansen i mine data er homogen mellom gruppene ved hjelp av to varianter av Levenes test.

3.4.3 P-verdier, effektstørrelser og statistisk styrke

Hvis vi har store nok utvalgsstørrelser, vil selv små forskjeller mellom to grupper gi seg utslag i en lav *p*-verdi. Vi kan dermed forkaste nullhypotesen, og hevde at vi har et signifikant funn (Larson-Hall, 2010, s. 96). I mange undersøkelser er det ofte vanskelig å få store nok utvalg, og *p*-verdien blir høy selv om det egentlig er en forskjell i populasjonen. Larson-Hall (2010, s. 101f) foreslår at hvis man tror at null-hypotesen ikke er sann, og man vil unngå en type 2-

feil, bør man sette alfa-verdien høyere enn 0,05. I mitt tilfelle kunne jeg antatt at PC-gruppa skriver tekster med bedre kvalitet og satt *p*-verdien til 0,1, som Larson-Hall (2010, s. 102) foreslår. Dette er det samme som å se bare på den ene «halen» til fordelingskurven, men det er ikke vanlig hvis en ikke gjentar en tidligere studie og kan være ganske sikker på at dette er riktig (Hatch & Lazaraton, 1991, s. 231). Jeg har ikke gjort dette, men jeg har oppgitt eksakt *p*-verdi, for det er stor forskjell på en *p*-verdi på 0,049 og en *p*-verdi på 0,001, selv om begge fører til at vi beholder nullhypotesen (Larson-Hall, 2010, s. 103). Jeg har også oppgitt eksakt *p*-verdi for Mann-Whitney U-test, som Andy Field (2005) sier er mer nøyaktig enn å angi om *p*-verdien er over eller under et nivå. Dette gjelder for små utvalgsstørrelser.

I tillegg har jeg oppgitt effektstørrelser, som Larson-Hall (2010, s. 114) hevder er den viktigste parameteren av alle. Med effektstørrelsene får vi innsikt i størrelsen på forskjellen mellom gruppene, sier hun. Effektstørrelsen er ikke avhengig av utvalgsstørrelsen, og vi kan se om vi har noe som er viktig å gå videre med, eventuelt med et større utvalg. For *t*-testene har jeg brukt Cohens *d* (Cohen, 1988) for effektstørrelse. I Mann-Whitney U-test har jeg regnet prosentvis varians etter formelen Cohens $r = Z/\sqrt{N}$, fordi Cohens *d* ikke er vanlig å bruke for Mann-Whitney U-test (Fritz et al., 2012, s. 12; Larson-Hall, 2010, s. 378). For Cohens *d* har jeg brukt retningslinjene gitt av Cohen (1988, s. 40), hvor $d = 0{,}2$ er en liten effektstørrelse, $d = 0{,}5$ er en middels effektstørrelse, og $d = 0{,}8$ er en stor effektstørrelse. For Cohens *r* gjelder at $r = 0{,}10$ er en liten effektstørrelse, $r = 0{,}30$ er en middels effektstørrelse, og $r = 0{,}50$ er en stor effektstørrelse (Cohen, 1988, s. 79f). Field (2000) anbefaler at man beregner statistisk styrke før man gjør en statistisk test. Dataprogrammet G*Power (Faul, Erdfelder, Lang & Buchner, 2007) viser at vi ikke kan få signifikant resultat for Cohens $d < 0{,}935$ for uavhengige *t*-tester med de parameterne jeg har satt ($\alpha = 0{,}05$, $\beta = 0{,}2$, $n = 19$, like store grupper, to-hale-test). For Mann-Whitney U-test må vi opp i 0,96 (Cohens *d*) for å beholde $\beta = 0{,}2$, som er vanlig i statistiske tester (Lowie & Seton, 2013, s. 48). β angir hvor stor sannsynligheten er for at vi ikke finner en effekt i utvalget som faktisk finnes i populasjonen (type 2-feil). Den statistiske styrken, det vil si sjansen til å finne en effekt som faktisk eksisterer i populasjonen, er altså ganske lav i min undersøkelse. Utvalgsstørrelsen skulle vært på 64 i hver gruppe for å finne en medium effektstørrelse (Cohens $d = 0{,}5$).

Jeg har satt alfa-verdien til 0,05 og oppgitt eksakte *p*-verdier. Jeg har også oppgitt effektstørrelser, Cohens *d* for t-test og Cohens *r* for Mann-Whitney U-test. Jeg har gjort korreksjoner for mange tester, undersøkt premissene for t-test og Mann-Whitney U-test og oppgitt eksakte *p*-verdier og effektstørrelser. Jeg har også undersøkt statistisk styrke og funnet

at min undersøkelse har for få deltakere til at jeg kan oppdage medium eller små effektstørrelser.

3.5 Oppsummering av metode

Jeg laget en prøve som liknet mest mulig på Norskprøven. Den ble gjennomført av 47 elever i to ulike grupper i to moduser: på PC og med penn på papir. Jeg har gjort rede for at deltakernes bakgrunn sannsynligvis ikke har betydning for undersøkelsen, og at gruppene er balanserte med hensyn til disse bakgrunnsvariablene. Etter en nivåvurdering av besvarelsene ble det 19 elever igjen i hver gruppe. De håndskrevne tekstene ble digitalisert og to mål ble tatt på alle besvarelsene. Jeg telte antall feil i besvarelsene og korrigerte dem. Deretter ble de andre variablene regnet ut. Jeg har brukt tre variabler for syntaktisk kvalitet, seks variabler for ordforråd og ett kombinert mål. Jeg har også anvendt noen statistiske metoder og gjort noen valg i forbindelse med dem. Det forskningsetiske i prosjektet er ivaretatt på en forsvarlig måte. Jeg skal nå gå over til å presentere resultatene av undersøkelsen.

4. Resultater

Jeg hadde fire hypoteser for å undersøke skriftlige tekster i to skrivemoduser. I disse hypotesene antok jeg at kvaliteten i tekstene øker når deltakerne skriver på PC og at kvaliteten påvirkes av alder og erfaring med bruk av PC som skriveverktøy. Jeg antok også at elever med tilstrekkelige PC-ferdigheter skriver bedre tekster på prøver med tidsbegrensning fordi de har mer tid til planlegging og revisjon når selve transkriberingen tar kortere tid. I den siste hypotesen antok jeg at konstruktvalideringen ved de nye norskprøvene ikke hadde vært god nok. Jeg brukte et utvalg syntaktiske og leksikalske variabler og analyserte dem statistisk. Jeg vil presentere resultatene av de statistiske testene, først for de syntaktiske variablene, deretter de leksikalske variablene, med oppsummering til slutt. Mål for hele ord er avrundet til nærmeste hele ord.

4.1 Syntaktiske variabler

4.1.1 Gjennomsnittlig lengde av t-enhet

Tabell 2. Gjennomsnittlig lengde av t-enhet

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	3,54	5,98	8,75	1,66
PC	19	3,54	5,45	7,31	1,23

Visuell inspeksjon og numerisk kontroll viste at data var normalfordelt, selv om begge grupper var platykurtiske, men innenfor anbefalt nivå (tabell 2). En Levenes test viste at variansen var homogen mellom gruppene ($F \approx 1{,}985$, $p = 0{,}167$). En uparet t-test viste ingen signifikans, $t(36) = 1{,}195$, $p = 0{,}240$. Effektstørrelsen var Cohens $d \approx 0{,}37$, altså en svak effekt. Resultatet viser at penn og papir-gruppa gjennomsnittlig skrev litt lengre t-enheter enn PC-gruppa, men at PC-gruppa hadde litt mindre spredning. Vi kan likevel ikke anta at dette gjelder for populasjonen.

4.1.2 Antall feil per t-enhet

Tabell 3. Antall feil per t-enhet

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	0,92	2,43	3,82	0,98
PC	19	0,83	2,75	4,06	0,93

Visuell inspeksjon og numerisk kontroll viste at data var normalfordelt, selv om penn og papir-gruppa var platykurtisk, dog innenfor anbefalte rammer (tabell 3). En Levenes test viste at variansen var homogen mellom gruppene ($F \approx 0{,}547$, $p = 0{,}464$). En uparet t-test viste ingen signifikans, $t(36) = 1{,}050$, $p = 0{,}301$. Effektstørrelsen var 0,34 (Cohens *d),* som betyr en svak effekt. PC-gruppa hadde litt flere feil, men litt mindre spredning enn penn og papir-gruppa. Dette kan ikke antas å gjelde for populasjonen.

4.1.3 Tekstlengde

Tabell 4. Tekstlengde i antall ord

Modus	N	Min	Middelverdi	Median	Maks	sd
Penn og papir	19	320	426	428	572	78,95
PC	19	269	433	452	762	112.41

Visuell inspeksjon og numerisk kontroll viste at data ikke var normalfordelt, PC-gruppa hadde en utligger, og var venstreskjev og leptokurtisk (tabell 4). En ikke-parametrisk Levenes test viste at variansen var homogen mellom gruppene ($F(1{,}36) \approx 0{,}084$, $p = 0{,}774$). En Mann-Whitney U-test viste at medianen var høyere hos PC-gruppa, og spredningen var størst i PC-gruppa. Likevel var det en tendens til at tekstene var høyere hos penn og papir-gruppa, for gjennomsnittlig rang var høyere hos penn og papir-gruppa, $U = 177$, $p = 0{,}931$ (Field, 2000, s. 52). Effektstørrelsen var Cohens $r \approx 0{,}017$, altså en svært liten effekt. Forskjellen var ikke signifikant.

4.1.4 Oppsummering syntaktiske variabler

Penn og papir-gruppa skrev altså gjennomsnittlig litt lengre t-enheter, men hadde litt større forskjell seg imellom enn PC-gruppa. PC-gruppa gjorde noe flere feil per t-enhet, og hadde litt lavere spredning enn penn og papir-gruppa. Penn og papir-gruppa hadde også noe lengre tekster, men medianen var høyere for PC-gruppa. Ingen av forskjellene var signifikante, så forskjellene gjelder bare for de elevene jeg har testet. GTEL og FTE hadde en svak effekt, Cohens *d* var over 0,3.

4.2 Leksikalske variabler

4.2.1 Gjennomsnittlig ordlengde

Tabell 5. Gjennomsnittlig ordlengde i antall tegn

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	2,79	3,36	3,98	0,39
PC	19	2,56	3,25	4,13	0,39

Visuell inspeksjon og numerisk kontroll viste at data var normalfordelt, selv om penn og papir-gruppa var noe platykurtisk (tabell 5), men innenfor anbefalt nivå, og en Levenes test viste at variansen mellom gruppene var homogen ($F \approx 0{,}571$, $p = 0{,}455$). En uparet t-test viste ingen signifikans, $t(36) = 0{,}874$, $p = 0{,}455$. Effektstørrelsen var Cohens $d \approx 0{,}29$, en liten effekt. Penn og papir-gruppa skrev gjennomsnittlig litt lengre ord, og hadde litt mindre spredning enn PC-gruppa. Forskjellen kan ikke infereres til populasjonen.

4.2.2 Malvern og Richards' D (vocd D)

Tabell 6. Tekstlengde på oppgave 2 i antall ord

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	112	163	253	38,47
PC	19	115	182	310	53,03

Malvern og Richards' D (vocd D) kunne, som nevnt, brukes på tekster mellom 100 og 400 ord. Ingen av mine tekster i oppgave 2 kom utenfor denne rammen, som tabell 6 viser.

Tabell 7. Malvern og Richards' D

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	49,22	53,95	60,84	3,95
PC	19	49,31	54,10	61,64	3,54

Visuell inspeksjon og numerisk kontroll viste at data var normalfordelt, men igjen var penn og papir-gruppa platykurtisk, men innenfor anbefalt nivå (tabell 7). En Levenes test viste at variansen mellom gruppene var homogen ($F \approx 0{,}997$, $p = 0{,}325$). En uparet t-test viste at forskjellen ikke var signifikant, $t(36) = 0{,}119$, $p = 0{,}906$. Effektstørrelsen var Cohens $d \approx 0{,}04$, en svært liten effekt. PC-gruppa lå så vidt over penn og papir-gruppa, og hadde litt større spredning. Dette gjelder ikke nødvendigvis for populasjonen.

4.2.3 Measure of textual lexical diversity

Measure of textual lexical diversity (MTLD) kunne ikke brukes på tekster under 100 ord, som tidligere nevnt. Jeg har vist i forrige mål at tekstene i oppgave 2 ikke var under dette.

Tabell 8. Measure of textual lexical diversity

Modus	N	Min	Middelverdi	Median	Maks	sd
Penn og papir	19	43,97	48,92	48,78	56,03	3,91
PC	19	44,67	48,42	47,87	56,12	3,21

Visuell inspeksjon og numerisk kontroll viste at data ikke var normalfordelt, PC-gruppa var høyreskjev (tabell 8). Penn og papir-gruppa var platykurtisk, men ikke utenfor anbefalt nivå. En ikke-parametrisk Levenes test viste at variansen mellom gruppene var homogen ($F(1,36) \approx 3{,}731$, $p = 0{,}061$). En Mann-Whitney U-test viste at MTLD var lavere hos PC-gruppa enn hos penn og papir-gruppa, $U = 175{,}00$, $p = 0{,}885$. Her viste alle sentralmålene samme tendens. Forskjellen var ikke signifikant. Effektstørrelsen var Cohens $r \approx 0{,}03$, altså svært nær null effekt.

4.2.4 Ordvariasjonsforhold

Jeg nevnte tidligere at variabelen ordvariasjonsforhold (OVF) kunne brukes på tekster mellom 85 og 170 ord. Tabell 9 viser at penn og papir-gruppa i oppgave 2 skrev tekster på mellom 112 ord og 253 ord, med en middelverdi på 163 ord (sd ≈ 38,47, N =19), mens PC-gruppa skrev tekster fra 115 ord til 310 ord, middelverdien var 182 ord (sd ≈ 53,03, N = 19). Åtte tekster måtte dermed fjernes fra hver av gruppene fordi de var for lange, og det ble 11 tekster igjen i hver gruppe.

Tabell 9. Ordvariasjonsforhold

Modus	N	Min	Middelverdi	Median	Maks	sd
Penn og papir	11	84,31	85,89	85,19	90,10	1,79
PC	11	81,63	86,57	85,56	91,43	2,96

Visuell inspeksjon og numerisk kontroll viste at data ikke var normalfordelt, penn og papir-gruppa var høyreskjev og leptokurtisk (men innenfor anbefalt nivå) og hadde en høy verdi på OVF = 90,10 (tabell 9). En ikke-parametrisk Levenes test viste at variansen mellom gruppene var homogen ($F(1, 20) \approx 1{,}274$, $p = 0{,}272$). En Mann-Whitney U-test viste at penn og papir-gruppa hadde lavere ordvariasjonsforhold enn PC-gruppa, $U = 50{,}50$, $p = 0{,}519$. Forskjellen var ikke signifikant, men effektstørrelsen var Cohens $r \approx 0{,}14$, altså en liten effekt. Dette kan ikke infereres til populasjonen.

Det viste seg at begrensningen oppad på 170 ord gjorde at åtte tekster i hver gruppe måtte fjernes fordi de var for lange, og den opprinnelige middelverdien for PC-gruppa lå høyere enn øvre grense for variabelen. Når en fjerner nesten halvparten av tekstene i den ene enden av fordelingen, mister variabelen mye av sin styrke.

4.2.5 Modifisert TTR

Modifisert TTR (MTTR) hadde ikke noen uttalt begrensning i tekstlengde, selv om den ikke er upåvirket av tekstlengde.

Tabell 10. Modifisert TTR

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	4,13	4,73	5,34	0,42
PC	19	4,09	4,73	5,32	0,40

Visuell inspeksjon og numerisk kontroll viste at data var normalfordelt, selv om begge grupper var platykurtiske, men innenfor anbefalt nivå (tabell 10). En Levenes test viste at variansen mellom gruppene var homogen ($F \approx 0{,}114$, $p = 0{,}737$). En uparet t-test viste at forskjellen mellom gruppene ikke var signifikant, $t(36) = 0{,}008$, $p = 0{,}994$. Effektstørrelsen var Cohens $d = 0$. MTTR for de to gruppene var lik. Dette er ikke nødvendigvis tilfelle i populasjonen.

4.2.6 Leksikalsk tetthet

Tabell 11. Leksikalsk tetthet

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	37,39	41,71	46,23	2,90
PC	19	38,46	41,21	45,84	2,07

Visuell inspeksjon og numerisk kontroll viste at data var normalfordelt, selv om penn og papirgruppa var høyreskjev (innenfor anbefalte rammer) og hadde mange verdier i midten (tabell 11). En Levenes test viste at variansen var homogen mellom gruppene ($F \approx 3{,}072$, $p = 0{,}088$), og en uparet t-test viste at forskjellen mellom gruppene ikke var signifikant, $t(36) = 0{,}363$, $p = 0{,}719$. Effektstørrelsen var Cohens $d \approx 0{,}12$, en svært svak effekt. Leksikalsk tetthet for de to gruppene var nærmest lik. Dette kan ikke infereres til populasjonen.

4.2.7 Lesbarhetsindeks

Tabell 12. Lesbarhetsindeks

Modus	N	Min	Middelverdi	Maks	sd
Penn og papir	19	14,83	20,19	26,31	3,76
PC	19	14,79	19,69	25,30	3,00

Visuell inspeksjon og numerisk kontroll viste at data var normalfordelt. Penn og papir-gruppa var høyreskjev og leptokurtisk, men innenfor anbefalt nivå (tabell 12). En Levenes test viste at variansen mellom gruppene var homogen ($F \approx 2{,}567$, $p = 0{,}118$), og en uparet t-test viste at forskjellen mellom gruppene ikke var signifikant, $t(36) = 0{,}455$, $p = 0{,}652$. Effektstørrelsen var Cohens $d \approx 0{,}15$, en svært svak effekt. Funnet kan ikke generaliseres til populasjonen. Tekstene hos PC-gruppa hadde litt lavere LIKS, og litt mindre spredning, men forskjellen var liten.

4.2.8 Oppsummering leksikalske variabler

Penn og papir-gruppa skrev altså gjennomsnittlig litt lengre ord, og hadde litt mindre spredning enn PC-gruppa. For vocd D lå PC-gruppa så vidt over penn og papir-gruppa, og hadde litt større spredning. For MTLD lå også penn og papir-gruppa litt høyere, og hadde litt større spredning enn PC-gruppa. Variabelen OVF viste at penn og papir-gruppa hadde litt mindre variert ordforråd, og mindre spredning enn PC-gruppa. OVF egnet seg ikke så godt fordi en god del av tekstene i én ende av utvalget må fjernes. For MTTR var gruppene nærmest identiske. Det var de også for leksikalsk tetthet, selv om penn og papir-gruppa hadde litt høyere middelverdi, hadde de også litt større spredning. For LIKS, som ikke er et rent leksikalsk mål, hadde tekstene hos PC-gruppa litt lavere verdi, og litt mindre spredning. Ingen av variablene ga signifikant resultat, og det høyeste effektmålet var for GOL, Cohens $d = 0{,}29$, en liten effekt.

4.3 Oppsummering for begge typer variabler

Det er viktig å huske på at ingen av mine funn var signifikante, det vil si at funnene ikke kan generaliseres til populasjonen. Hovedgrunnen til dette er sannsynligvis at det er for få deltakere i undersøkelsen. Som jeg nevnte tidligere, måtte jeg hatt 64 deltakere i hver gruppe for å finne en middels effektstørrelse, og enda flere for å finne en liten effektstørrelse. Selvfølgelig kan det også være slik at variablene ikke er forskjellige mellom de to skrivemodusene. Hvis vi ser på effektstørrelsen for de enkelte variablene, finner vi at gjennomsnittlig t-enhetslengde, antall feil per t-enhet og gjennomsnittlig ordlengde var de eneste variablene som hadde over Cohens $d = 0{,}25$ i effektstørrelse. Det vil si at disse variablene var de som viste størst utslag mellom de to skrivemodusene. To av disse er syntaktiske, den siste variabelen er leksikalsk. På seks av variablene (GTEL, FTE, TL, GOL,

MTLD og LT) hadde penn og papir-gruppa høyere verdier, på tre andre variabler (vocd D, OVF og LIKS) var det PC-gruppa som var høyest, mens det var likt på den siste variabelen (MTTR).

4.4 Tekstlengde og feil hos de eldste deltakerne

I kapittel 2 argumenterte jeg for at eldre innvandrere kan ha mindre trening på PC enn de yngre. Jeg antok at det ville gi seg mest utslag i at skrivehastigheten var lav, og at total tekstlengde dermed ville være lavere enn for yngre deltakere. Videre trodde jeg at antall feil ville øke når deltakerne fikk dårlig tid. Jeg ville undersøke dette for deltakere som var 40 år eller eldre, det vil si at de antagelig har lært å bruke digitale verktøy i relativt voksen alder. Dette gjelder sannsynligvis i stor grad innvandrere, som jeg drøftet i kapittel 2.4.1. Både penn og papir-gruppa og PC-gruppa hadde hver fire deltakere som var over 39 år. I dette tilfellet ville en toveis ANOVA vært riktig statistisk test å bruke, men den forutsetter normalfordelte utvalg (Lowie & Seton, 2013, s. 63), og det er vanskelig å si om utvalg på fire i hver gruppe er normalfordelt. Toveis ANOVA skal også ha minst fem observasjoner i hver gruppe (celle) (Hatch & Lazaraton, 1991, s. 384). Et ikke-parametrisk alternativ finnes ikke, Kruskal-Wallis kan brukes for å teste hver uavhengig variabel separat, men vi får ikke vite noe om interaksjonen mellom variablene (Larson-Hall, 2010, s. 142). Dessuten har Kruskal-Wallis lite styrke, et totalt utvalg på sju eller lavere gir *p*-verdi over 0,05 uansett (GraphPad, 2016). I mitt tilfelle er det totale utvalget åtte. Jeg velger derfor å vise nøkkeltall for total tekstlengde og antall feil per t-enhet for de to skrivemodusene i tabellform (tabell 13 og 14).

Tabell 13. Total tekstlengde i antall ord for skrivemodus og alder

Modus	Alder	N	Min	Middelverdi	Median	Maks	sd
Penn og papir	Yngre	15	325	425,93	428	538	68,26
	Eldre	4	320	428,25	410,50	572	125,11
PC	Yngre	15	308	442	452	762	110,68
	Eldre	4	269	401,25	400,50	535	129,99

Vi ser at total tekstlengde (tabell 13) var en god del lavere hos de eldre i PC-gruppa enn hos de yngre i samme skrivemodus, medianen var 51,5 ord lavere for de eldre, og middelverdien var 40,75 ord lavere, men det var en svært høy verdi på 762 ord som dro opp de yngste mye.

Variasjonen var også stor. Effektstørrelsen for middelverdien mellom de yngste og de eldste i PC-gruppa var Cohens $d \approx 0{,}38$, en liten effekt, men gruppene var svært små, særlig gruppa med de eldste. Forskjellen mellom aldersgruppene var mindre for de som brukte penn og papir, der var middelverdien bare 2,32 ord lavere og medianen 17,5 ord lavere for de eldre, men variasjonen var stor her også, særlig for de eldre. Effektstørrelsen for middelverdien mellom de eldste og de yngste i penn og papir-gruppa var Cohens $d \approx 0{,}03$, altså ingen effekt. Igjen var gruppene små, særlig gruppen med de eldste, og disse tallene kan naturligvis ikke infereres til populasjonen.

Tabell 14. Feil per t-enhet for skrivemodus og alder

Modus	Alder	N	Min	Middelverdi	Median	Maks	sd
Penn og papir	Yngre	15	0,92	2,37	2,27	3,82	1,08
	Eldre	4	2,14	2,63	2,59	3,21	0,52
PC	Yngre	15	0,83	2,52	2,73	3,84	0,90
	Eldre	4	3,06	3,62	3,69	4,06	0,44

For feil per t-enhet (tabell 14) er det i begge moduser forskjell på antall feil i aldersgruppene i favør av de yngre, men forskjellen er klart større i PC-gruppa enn i penn og papir-gruppa. I PC-gruppa er det mer enn én feil per t-enhet i forskjell mellom aldersgruppene både i middelverdi og median. Effektstørrelsen for middelverdien hos penn og papir-gruppa var Cohens $d \approx 0{,}27$, en liten effekt. Effektstørrelsen for middelverdien hos PC-gruppa var Cohens $d \approx 1{,}39$, altså en stor effekt. Blant de yngre hadde begge skrivemoduser fire deltakere med mindre enn to feil per t-enhet, dette gjaldt ingen av de eldre. PC-gruppa hadde ingen under tre feil per t-enhet blant de eldre. Blant de yngre finner vi de som gjør flest feil, og de som gjør færrest feil, spredningen hos dem er større. Det er særlig stor forskjell på antall feil for PC-gruppa, de eldre gjør mange flere feil enn de yngre. Igjen må jeg ta forbehold om at gruppene er små, tallene gjelder sannsynligvis bare for utvalget.

Det var en liten effekt av skriveverktøyet på tekstlengde for de eldste av mine deltakere, men blant de yngste hadde én deltaker en svært høy verdi, og det var svært få deltakere i gruppa med de eldste deltakerne. Det var en større effekt av skriveverktøyet blant de eldste av deltakerne når det gjaldt antall feil per t-enhet. Ingen av funnene var signifikante. Dette fører meg over til en diskusjon av funnene i lys av teorier og tidligere forskning.

5. Diskusjon

Jeg formulerte fire hypoteser for å undersøke kvaliteten på skriftlige andrespråkstekster i to skrivemoduser. I den første hypotesen antok jeg at kvaliteten i skriftlige andrespråkstekster øker når deltakerne skriver på PC. I den andre hypotesen antok jeg at alder og erfaring med bruk av PC som skriveverktøy virker inn på tekstenes kvalitet. I den tredje hypotesen antok jeg at elever med tilstrekkelige PC-ferdigheter skriver bedre tekster på prøver med tidsbegrensning fordi de har mer tid til planlegging og revisjon når selve transkriberingen tar kortere tid. I den siste hypotesen antok jeg at konstruktvalideringen ved de nye norskprøvene ikke hadde vært god nok. Jeg brukte et utvalg syntaktiske og leksikalske variabler og analyserte dem statistisk.

Som jeg nevnte, hadde penn og papir-gruppa høyere verdier på seks av variablene (GTEL, FTE (altså færre feil), TL, GOL, MTLD og LT), mens PC-gruppa lå høyest på tre andre variabler (vocd D, OVF og LIKS), og det var likt på den siste variabelen (MTTR). Forskjellene var ganske små, og ikke signifikante. Likevel var det interessant at penn og papir-gruppa hadde litt høyere verdier for lengdemålene (unntatt LIKS, som også delvis er et lengdemål) og leksikalsk tetthet, mens PC-gruppa bare lå høyest på noen av variablene for leksikalsk variasjon. Andre mål for leksikalsk variasjon var høyest i penn og papir-gruppa. Det er mulig at PC-gruppa skrev kortere tekster fordi de ikke hadde nok erfaring med PC, tekstbehandling og tastatur, men det er mest sannsynlig at den kortere tekstlengden var tilfeldig. Likedan er det også mulig at PC-gruppa fikk litt flere feil fordi de ikke behersket teknologien tilstrekkelig, men igjen er det sannsynligvis en tilfeldighet at de var litt høyere på FTE. Når ikke alle variablene for leksikalsk variasjon lå høyest i én gruppe, kan det tyde på at det var tilfeldig. Selvfølgelig kan det også bety at én eller flere av variablene målte forskjellige konstrukter, eller at variablene ikke er brukbare til en slik undersøkelse som jeg har foretatt her.

Med funn som ikke var signifikante, er det vanskelig å bekrefte hypotesene jeg satte opp innledningsvis. Den første hypotesen antok at kvaliteten økte når deltakerne skrev på PC. Mine elever i PC-gruppa hadde tendenser til kortere t-enhetslengde, flere feil per t-enhet, kortere total tekstlengde, kortere gjennomsnittlig ordlengde og lavere leksikalsk tetthet. På variablene for leksikalsk variasjon var det forskjellige resultater, begge gruppene var høyere på noen variabler, mens det var likt på én variabel. Tidligere forskning jeg har referert (Bangert-Drowns, 1993; Collins mfl., 2013; Goldberg mfl., 2003; Pennington, 2003; Roblyer mfl., 1988; van Waes, 1994) fant at tekstenes kvantitet og kvalitet økte når elevene brukte PC i

stedet for penn og papir. For flere typer revisjon var funnene sprikende, antagelig fordi revisjon måles på ulike måter. De fleste av disse undersøkelsene har vært gjort på yngre elever enn i min undersøkelse, men Pennington (2003) så på voksne elever og gjorde de samme generelle funnene som nevnt over. Hun poengterte at holdninger til skriving på PC var viktig, slik at positive holdninger ville påvirke skrivingen og gi større lengde og bedre kvalitet, mens negative holdninger ville gi det motsatte. Både ferdigheter og holdninger kan naturligvis ha forandret seg siden Penningtons undersøkelse. Hos mine elever var det, som antydet tidligere, ikke noe som tyder på at kvalitet og kvantitet økte for de som skrev på PC, heller tvert imot. Tekstenes kvalitet målt med de tekstlige variablene jeg har anvendt forandret seg ikke mye mellom modusene, forskjellene var små. Igjen kan dette naturligvis være tilfeldig. Antall deltakere var for lite til å finne eventuelle små forskjeller, men likevel er det mulig å drøfte på bakgrunn av funnene mine og forskningen og teorien jeg har gjort rede for.

Vi så at særlig kompleksitetskonstruktet var komplisert og sannsynligvis sammensatt av flere konstrukter. I min undersøkelse var det ulike resultater for variablene som målte kompleksitet: både penn og papir- og PC-gruppa hadde best resultater på noen av variablene, og noen var like. Dette kan støtte hypotesen hos Larsen-Freeman (1997; 2009) om at andrespråkstilegnelse og variabler for kvalitet må forstås i lys av kompleksitetsteori. Et av problemene i diskusjonen om kvalitet i skriving er at både konstruktene og variablene defineres forskjellig av alle forskere, i tillegg til at metodene er ulike. Derfor blir det også krevende å sammenligne, og metastudier får liten verdi. Tall som er fremkommet på bakgrunn av ulike variabler og metoder kan vanskelig sammenlignes, selv om effektstørrelsene standardiseres. Dette kan også være en av grunnene til at kompleksitet framstår som så komplekst.

I den andre hypotesen min antok jeg at alder og erfaring med bruk av PC som skriveverktøy spilte en rolle for kvaliteten på tekstene. De eldste elevene i mitt materiale som brukte PC skrev noe kortere enn de yngre i samme skrivemodus. Det var en svak effekt av skriveverktøyet, for det var ingen forskjell mellom eldre og yngre i penn og papir-gruppa når det gjaldt tekstlengde. Den eldste gruppen besto av bare fire elever i hver modus, og resultatene var ikke signifikante, så vi vet ikke om dette var tilfeldig. Jeg fant også at det var stor forskjell på antall feil per t-enhet mellom aldersgruppene i PC-gruppa, i PC-gruppa var effekten av alder stor, mens i penn og papir-gruppa var det en liten alderseffekt. Igjen var gruppene små, det var bare fire elever i de eldste gruppene i hver modus, og funnene var ikke signifikante. Aldersforskjellene kan forklares med funnene jeg har nevnt over, om at eldre som ikke har tilstrekkelige ferdigheter med PC og tekstbehandling, er forfordelt når en tester skriving på PC

(Dunn & Reay, 1989; Horkay et al., 2006; Madsen, 1991; McNamara, 2000; Neu & Scarcella, 1991; Pennington, 2003; Rusmin, 1999; van Dijk, 2006; Warschauer & Liaw, 2010; Wolfe & Manalo, 2004). Noen av tekstene i mitt materiale var preget av hastverk, og noen elever manglet svar på oppgave 3, dette kan forklares med at tiden var en avgjørende faktor for noen grupper (Hamp-Lyons & Kroll, 1996; Murphy & Yancey, 2008). Det er plausibelt at antall feil påvirkes av skrivehastighet, for hvis en kandidat ser at det er mye igjen å gjøre og tiden renner ut, er det sannsynlig at vedkommende prioriterer å skrive seg ferdig framfor å rette opp feil og lese over det som er gjort. Igjen må jeg ta forbehold for små grupper.

Den tredje hypotesen stilte spørsmål om elever som hadde tilstrekkelige PC-ferdigheter skriver bedre tekster fordi de hadde mer tid til planlegging og revisjon når selve transkriberingen tar kortere tid og om dette gjelder for andrespråkselever. Bangert-Drowns (1993) hevdet at dette var tilfelle for førstespråkselever og Pennington (2003) fant det samme for andrespråkselever. Det er vanskelig å vite hvilke prioriteringer elevene har gjort, noen har kanskje brukt tid på å planlegge og revidere, mens andre har skrevet langt. Vi så at gruppene som helhet ikke var særlig forskjellige, seks mål var bedre hos penn og papir-gruppa, mens tre mål var bedre hos PC-gruppa, og for ett mål var det likt. Når vi ser på de yngre i PC-gruppa alene, ser vi noe tydeligere forskjell, i PC-gruppa hadde de yngre en liten tendens til å skrive lengre tekster enn de eldre. For feil per t-enhet var det større forskjell, de yngre i PC-gruppa hadde klart færre feil enn de eldre i samme gruppe. Vi kan anta at de yngre hadde større skrivehastighet, og har utnyttet dette til å få færre feil da de nådde anbefalt ramme for tekstlengde gitt i oppgaven. Det hadde vært interessant å se på flere variabler, for eksempel gjennomsnittlig t-enhetslengde eller gjennomsnittlig ordlengde, for å finne om de yngre brukte tiden til å øke også dette. I denne avhandlingen hadde jeg ikke kapasitet til det. Som nevnt, er det viktig å huske at gruppene var små, det gjaldt særlig de eldre.

Den siste hypotesen jeg undersøkte stilte spørsmål om konstruktvalideringen ved de nye norskprøvene har vært god nok. Vi så at mine deltakere i PC-gruppa hadde en liten tendens til å skrive kortere tekster og til å ha flere feil enn deltakerne i penn og papir-gruppa. Vi så også at de eldre deltakerne som skrev på PC hadde en liten tendens til å skrive kortere tekster og en tydelig tendens til flere feil enn de yngre i samme skrivemodus. Messick (1989) kalte den ene typen trusler mot konstruktvaliditeten for konstrukt-irrelevant varians. Han mente at noe hadde kommet inn i testen som ikke tilhørte konstruktet. Hvis dette gjorde testen vanskeligere, kalte han det «construct-irrelevant difficulty». Det er rimelig å anta at eldre andrespråkselever med svake ferdigheter på PC og med tekstbehandling opplever slik vanskelighet når de testes i

tidsbegrenset skriving på PC (Wolfe & Manalo, 2005). Jeg har også gjort rede for at den som konstruerer prøver skal undersøke om visse grupper vil bli påvirket av prøven. Jeg vet ikke om Vox har gjort dette, men jeg har ikke funnet noe publisert om det.

Med bakgrunn i det jeg nå vet, ville jeg kanskje brukt en etablert metode for analytisk vurdering av tekster, for eksempel den såkalte *Jacobs' scale* (også kalt *ESL Composition Profile*), som vurderer fem tekstlige nivåer: innhold, organisering, vokabular, språkbruk og ortografi/tegnsetting (mechanics). Disse er vektet i den rekkefølgen de er nevnt her (Weigle, 2002). Bruk av denne skalaen hadde gjort det enklere å sammenligne med annen forskning, med forbehold om at norsk og engelsk er forskjellig. Jacobs' scale likner vurderingsmatriser som brukes i ungdomsskolen og gir kanskje en mer helhetlig vurdering av tekstene enn variabler for kompleksitet, nøyaktighet og flyt. Ulempen er at den (som Norskprøven) i stor grad er avhengig av vurderinger og flere personer til å vurdere, og at disse trenger trening i å bruke skalaen.

Jeg skulle også gjort en pilotstudie for å finne ut om metoden og variablene jeg tenkte å bruke var anvendelige i denne typen forskning (Mackey & Gass, 2005, s. 158). Det hadde også gitt mer styrke om jeg hadde brukt de samme elevene i begge modus, som jeg har nevnt. Det var praktiske årsaker som forhindret dette. Statistisk burde jeg kanskje gjort en prinsipalkomponentanalyse og deretter en multippel regresjonsanalyse for å finne hvilke variabler som bidro mest til konstruktene. Da hadde analysen blitt tydeligere, og muligens gitt mer styrke. Selvfølgelig skulle jeg også hatt flere deltakere. Det var ikke mulig kapasitetsmessig.

Jeg ville også gjerne gå inn og se mer på enkeltelever og –tekster for å se hva de har tenkt og hvordan de har formet tekstene sine. En casestudie eller annen kvalitativ synsvinkel (for eksempel såkalt *think-aloud protocol*) hadde vært interessant i den sammenhengen. Når alt dette er nevnt, hadde jeg likevel interessante funn. Jeg fant at penn og papir-gruppa fikk best uttelling på lengdevariablene, variabelen for feil og to av de leksikalske variablene, mens PC-gruppa fikk bedre verdier på to leksikalske variabler og den kombinerte variabelen, mens det var lik uttelling for den siste leksikalske variabelen. Jeg fant høyest effektstørrelse ved å se på forskjellen mellom de eldste i hver skrivemodus. Særlig var det tydelig flere feil blant de eldste som skrev på PC, men også for lengde var det en liten forskjell i mitt materiale i favør av de som skrev med penn på papir blant de eldste.

Som jeg nevnte i innledningen, finnes det naturligvis flere årsaker til at resultatene ble svakere da Norskprøven ble digital. Det kom samtidig inn en ny oppgavetype, å skrive til bilde, som kan gi et enklere språk av typen «Jeg ser…», altså fortellende helsetninger. For nivået B1 skal kandidaten ha «…noe variasjon i setningstyper. Også noe vellykket bruk av komplekse setninger» som det står i vurderingsskjemaet som sensorene bruker (Vox, 2016g). Kandidaten skal også vise kohesjon i teksten. Selv om sensorene skal vurdere alle tre oppgavene samlet, kan et svært enkelt språk i oppgave 1 («Skriv til bildet») gi sensorene et førsteinntrykk som er negativt. I denne sammenhengen nevner Wolfe & Manalo (2005) noe annet som gi sensorene et feilaktig inntrykk: en tekst som er skrevet på PC ser kortere ut enn en som er skrevet med penn og papir. Sensorene har ofte lang erfaring og har sett mange papirbaserte prøvetekster, og de reagerte kanskje negativt, bevisst eller ubevisst, da de fikk se den første runden med digitale tester. Som nevnt, har resultatene på prøvene tatt seg opp etter hvert som den nye norskprøven har blitt gjennomført noen ganger, noe som kan bekrefte denne antagelsen.

Endelig skal vi ikke glemme at det ble åpnet for at elevene kunne melde seg på Norskprøven selv da den ble digital, vi fikk altså en ny populasjon som meldte seg på. Dette gjorde nok at en del av elevene ikke brukte læreren som «filter», noen hadde kanskje et urealistisk bilde av sine egne norskferdigheter. Vox har en såkalt «Nivåvelger» på sine nettsider (Vox. 2016b), der elevene kan krysse av for hva de kan og få en anbefaling for hvilket nivå de skal melde seg opp til. Denne er ganske rundt og utydelig formulert: «Jeg kan skrive sammenhengende tekster» og «Jeg kan skrive tekster som er lette å forstå» er ikke helt entydige og lette å operasjonalisere for elever som ikke har mye trening i å vurdere seg selv og kanskje ikke har fått tekster vurdert av en lærer. Nivåvelgeren nevner heller ikke at PC-ferdigheter eller skrivehastighet kan ha en betydning for oppnådd nivå. Noen elever kan også ha fått gode tilbakemeldinger på sitt muntlige nivå og tror det også gjelder det skriftlige. Skriftlig produksjon er mest krevende for elevene, og det er færrest som oppnår B1-nivå her (Vox, 2016d). Hvis elevene trenger eller ønsker B1-nivå på alle ferdighetene og ikke kjenner sitt eget nivå, må de kanskje ta ny skriftlig prøve senere. De vil uansett få en nivåplassering, de kan få B1, A2, A1 under A1, eller «ikke vurdert».

Det var altså vanskelig å få signifikante funn og dermed å bekrefte hypotesene i mitt prosjekt, noe som kan være tilfeldig, men som også kan skyldes at antall deltakere i undersøkelsen var for lite. I den første hypotesen antok jeg at kvaliteten i skriftlige andrespråkstekster øker når deltakerne skriver på PC, men funnene pekte ikke entydig i den retningen. For den andre hypotesen, som tilsa at alder og erfaring med PC virker inn på tekstenes kvalitet, var det

tydeligere resultater, men heller ikke disse var signifikante. I den tredje hypotesen antok jeg elever med tilstrekkelige PC-ferdigheter skriver bedre tekster på prøver med tidsbegrensning fordi de har mer tid til planlegging og revisjon når transkriberingen tar kortere tid. De yngre i min undersøkelse skrev noe lengre tekster, men særlig brukte de tiden til å få færre feil. Også her manglet signifikansen. I den siste hypotesen antok jeg at konstruktvalideringen av de nye norskprøvene ikke hadde vært god nok. Hvis vi skal dømme ut fra elevene som deltok i mitt prosjekt, har den ikke vært tilstrekkelig. Det er mulig at Norskprøven inneholder konstrukt-irrelevant varians når den gjennomføres på PC av elever med svake PC-ferdigheter. Igjen er det viktig å huske at resultatene ikke var signifikante.

I ettertankens lys burde jeg ha brukt en etablert skala for å måle kvalitet i tekstene, gjort en pilotstudie og mer avanserte statistiske analyser. Jeg skulle også gjerne ha gjort mer kvalitative studier for å få et mer helhetlig bilde av elevenes virkelighet. Det er også sannsynlig at andre årsaker ligger bak de svakere resultatene da Norskprøven ble digital. Også andre deler av prøven ble endret samtidig, sensorene kan ha blitt påvirket, og en ny populasjon meldte seg på prøven, kanskje uten den samme veiledningen fra en lærer. Veiledningen på Vox' hjemmesider er også litt utydelig og vanskelig å bruke som egenvurdering for denne elevgruppen, blant annet mangler en referanse til digitale ferdigheter.

Jeg skal nå antyde noen implikasjoner fra dette prosjektet, først noen som gjelder videre forskning, deretter noen didaktiske implikasjoner og til slutt noen anbefalinger for testing. Naturligvis vil noen av disse gå noe over i hverandre, for utgangspunktet for de tre er det samme.

5.1 Forskningsmessige implikasjoner

Som nevnt, var det vanskelig å få signifikante resultater. Med flere deltakere i hver gruppe hadde det vært lettere å se om forskjellene i min undersøkelse også finnes i populasjonen. Larson-Hall (2011, s. 249) hevder at det er sannsynlig at funnene vil bli signifikante med flere deltakere hvis effektstørrelsen er stor. I noen av mine resultater var effektstørrelsen relativt stor, selv om det naturligvis kunne vært tilfeldig, så det hadde vært interessant å gjenta undersøkelsen med flere deltakere. For rettferdig behandling av kandidater til Norskprøven er det viktig å undersøke grundig om validiteten til prøven er god nok, som jeg drøftet i kapittelet om rettferdighet. Særlig er det viktig å se på konstrukt-irrelevant varians, at elever med svake PC-ferdigheter ikke klarer den delen av Norskprøven som måler skriftlig produksjon nettopp

fordi de ikke skriver raskt nok på PC. Det gjelder kanskje ikke mange kandidater, men det er viktig for dem det gjelder (Wolfe & Manalo, 2005, s. 7), slik som jeg har funnet tendenser til i mitt prosjekt. Jeg anslo tidligere at mellom 1000 og 2500 eldre tar Norskprøven hvert år (se kap. 2.4.1). Hvis vi videre antar at 10 % av disse har svake PC-ferdigheter, vil det være mellom 100 og 250 personer som har problemer med skriftlig produksjon hvert år fordi de ikke behersker verktøyet i tilstrekkelig grad. Naturligvis er disse tallene bare antagelser, men allikevel er det altså alvorlig for de personene det gjelder. For å unngå urettferdig behandling av disse elevene, bør en vurdere å gjøre en fornyet konstruktvalidering av Norskprøven når den tas digitalt av eldre kandidater. Konstruktvalidering er en forpliktelse Vox har etter de etiske retningslinjene i ALTEs praksiskodeks (ALTE, 2007). Som jeg har nevnt, vil også casestudier eller intervjuer med lærere og elever kunne gi innsikt i situasjonen for dem som er berørt.

Jeg har tidligere vist at mange måleenheter og metoder har vært i bruk for å måle kompleksitet, nøyaktighet og flyt, og at særlig kompleksitet kan bestå av flere konstrukter. Hvis konstruktene er ulike, har de svak *konvergensvaliditet* (Messick, 1989, s. 47). I mitt prosjekt var det sprikende resultater fra de ulike kompleksitetsvariablene, noe som kan peke i retning av at det er flere konstrukter. Gabriele Pallotti (2009, s. 599) oppfordrer forskere til å komme til en konsensus om konstrukter og en operasjonalisering av dem, slik at det blir mulig å sammenligne resultater på tvers av studier. Hun hevder at konstruktene kan standardiseres, men minner også om at det er vanskelig å måle alle fasetter ved kompleksitet, nøyaktighet og flyt i språket. Dette berører det evige problemet i språktesting: at ferdigheten selv brukes til å teste ferdigheten (Pedersen, 2008, s. 98).

Etter hvert som det blir avlagt et antall prøver i det digitale formatet, bør det også forskes på autentiske tekster fra Norskprøven, både kvantitativt og kvalitativt, for eksempel gjennom korpuslingvistikk. Da kan man se om det tegner seg et bilde som ligner det jeg har funnet. Norsk Språktest hevder at de digitale og de papirbaserte prøvene ikke kan sammenlignes, men det forhindrer ikke at det kan forskes på de digitale besvarelsene alene, og at man kan undersøke om eldre kandidater med svakere PC-ferdigheter har en ulempe i forhold til de yngre når de bruker tekstbehandling.

I kapittel 2.3 var jeg inne på at det er lite forskning på andrespråksskriving i Norge. Dette gjelder ikke bare for sammenligning av ulike moduser, men også for andre typer klasseromsforskning som fokuserer på arbeid med digitale ferdigheter. Jeg refererte tidligere

funn fra Horkay et al. (2006) som viste at barn og ungdom med svake ferdigheter i tekstbehandling har en ulempe i skriving, selv om de bruker PC mye, både hjemme og på skolen. Vi tar det kanskje for gitt at dagens barn og unge kan å bruke PC, men det gjelder ikke alltid tekstbehandling. I voksenopplæring ble digital kompetanse innført som en basiskompetanse integrert i de språklige kompetansemålene med den nye læreplanen fra Vox som kom i 2012. Det er nå fire år siden, og den nye læreplanen bør evalueres etter at den har vært i bruk en stund. Også i andre skoleslag bør det forskes mer på andrespråksskriving når vi vet at antall andrespråkselever er økende, og fortsatt vil øke mye framover. Vi har også sett at skriving er det mest krevende for kandidatene på Norskprøven.

Det er også gjort for lite forskning på hvorvidt tilgang på og erfaring med PC påvirker resultatene når andrespråkselever skal skrive lengre tekst på PC, sier Wolfe og Manalo (2005, s. 6). De snakker om internasjonal forskning, men det gjelder i enda større grad i Norge. Jeg har ikke funnet noe norsk (eller skandinavisk) forskning på voksne andrespråkselever som skriver på PC. Som jeg har vist er det vanskelig å inferere funn fra morsmålselever og yngre elever til å gjelde eldre andrespråkselever. Det bør gjennomføres et større prosjekt som ser på dette, med tilstrekkelig antall deltakere slik at man finner forskjeller som kan antas å være mellom gruppene, og som rapporterer effektstørrelser. Med det jeg har vist av potensiell konstrukt-irrelevant varians, bør det være en svært aktuell tilnærming.

5.2 Didaktiske implikasjoner

Jeg tror ikke at voksenopplæringssentrene og lærerne der er klar over at elever med svake ferdigheter på PC kan ha en såpass stor ulempe i skriftlig framstilling som jeg har funnet tendenser til, og som jeg har referert fra ulike forskere i dette prosjektet. Det er videre sannsynlig at disse institusjonene og lærerne ikke i tilstrekkelig grad tar dette innover seg i praksis. De bør gjøre elevene klar over at tekstlengde er et kriterium som sensorer, mer eller mindre ubevisst, kan legge vekt på i vurderingen av besvarelsen. Elevene bør også være klar over at sensorer kan stille høyere krav til en tekst skrevet på PC enn en tekst skrevet med penn og papir, og at en tekst på PC kan framstå som kortere, som Wolfe og Manalo (2005, s. 10) refererer. I kapittel 3.3.1 om tekstlengde over har jeg vist at tekstlengde og karakter korrelerer, både i første- og andrespråksskriving. Likevel er det ikke sånn at tekstlengde alene gir bedre karakter, karakteren og tekstlengden reflekterer sannsynligvis skriveferdigheter også på andre

tekstlige nivåer. Iallfall bør elevene få opp skrivehastigheten på PC, hvis andre deler av skriveferdighetene er tilstrekkelige.

Da de nye norskprøvene skulle innføres, fikk lærerne en innføring i prøvene fra Vox. Denne innføringen fokuserte ikke på skriveferdigheter på PC, men på at skolene måtte ha tilstrekkelig utstyr. Den slo også fast at prøven i skriftlig framstilling var det vanskeligste for elevene, men problematiserte ikke dette noe mer og ga ingen anbefalinger om trening av digitale ferdigheter (Vollan, 2014). Det er mulig at Vox ikke var klar over at svake PC-ferdigheter hos noen av elevene ville slå ut, men det burde vært gjort tydeligere for lærere og kandidater at de må trene elevene mer i å skrive raskt og bruke PC på den mest nyttige måten. Iallfall burde Vox informert lærere og kandidater bedre om kravene til undervisning i digital skriving og at et visst nivå på PC-ferdigheter var nødvendig da de første resultatene av den nye norskprøven viste en kraftig reduksjon i antall bestått prøver i skriftlig framstilling (fra ca. 50 % til ca. 30 %).

I tillegg bør Vox endre den såkalte «Nivåvelgeren» på sin nettside (Vox, 2016b), der kandidatene kan vurdere sitt eget nivå, slik at den reflekterer kravet til digital skriving. Slik er det ikke nå, den nevner ikke digitale ferdigheter i det hele tatt. Det bør også angis minimum skrivehastighet, for eksempel fem ord i minuttet. Mine elever, som var over A2-nivå, skrev mellom 2,99 og 8,47 ord i minuttet, middelverdien lå på 4,82 (sd ≈ 1,25). I tillegg var det noen elever som ikke fullførte oppgave 3, muligens fordi de skrev for sakte. Jeg har regnet tallene ut fra at de skriver i 90 minutter, men de må også bruke tid til å lese oppgaven, planlegge og revidere teksten, så egentlig er selve skrivehastigheten høyere. I dag angis det ingen slike krav på «Nivåvelgeren» hos Vox, denne siden nevner faktisk ikke at PC brukes i det hele tatt, selv om det står andre steder på nettstedet.

MacArthur (2006), Pennington (2003) og flere andre forskere viser at elevene må få undervisning i bruk av PC og tekstbehandling for at tekstene på PC skal bli bedre enn håndskrevne tekster. Bangert-Drowns (1993) sier at elevene må få undervisning i *det som gir PC en fordel* framfor penn og papir. Jeg tror lærerne må gå mer detaljert til verks enn i dag, og vise elevene hvordan de bruker rettefunksjon, klipp-og-lim, flytting, søking, navigasjon, overskriftssstiler og alle de andre mulighetene som ligger i en PC. Dette læres antagelig ikke uten eksplisitt instruksjon eller at elevene ser at de har nytte av det, og mange av disse funksjonene er vanskelig tilgjengelige, eller krever at man utfører flere prosesser. Elevene får ikke nødvendigvis bruk for alt dette på Norskprøven (slik den er i dag), men de får bruk for

det i utdanningen og i senere arbeidsforhold. Også høy skrivehastighet kan gi skriving på PC en fordel framfor penn og papir, som jeg fant i prosjektet. Dette må også trenes opp, for eksempel ved at elevene lærer den såkalte «touch-metoden» som gjør at de slipper å flytte blikket ned på tastaturet mens de skriver og automatiserer skriving i større grad ved tilstrekkelig trening. Elevene må få forståelsen av det er mengdetrening som skal til, og at de må jobbe med dette også utenfor skolen. For trening hjelper, sier Wolfe og Manalo (2005, s. 8), som refererer undersøkelser hvor eldre kandidater økte sine resultater i skriving betydelig ved å trene på bruk av PC.

Jeg har også vært inne på at andrespråkselever har utfordringer når det gjelder skriveprosessen (Magnusson, 2013; Reynolds, 2005; Silva, 1993). Kanskje må lærerne bryte skriveprosessen ned i sine enkelte bestanddeler og hjelpe elevene med å utføre dem. For eksempel kan de vise hvordan oppgaven tolkes, hvordan man noterer idéer, hvordan man planlegger, strukturerer besvarelsen og reviderer, både lokalt og globalt, i enda større grad enn i dag. Prosesskriving er en nyttig måte å gjøre dette på, da tar man tak i den enkelte elevs styrker og svakheter og gir underveisvurderinger som kan løfte elevens skriveferdigheter der det er mest nødvendig. Hvis elevene i tillegg får forståelse for at alle disse oppgavene kan gjøres enklere på en PC, og at de er nyttige senere, kan motivasjonen øke (Pennington, 2003). Man bør også bruke skriverammer og modelltekster for å vise hvordan en tekst bør bygges opp for de enkelte sjangre. Problemet er naturligvis at alt dette tar tid, at elevene skal lære mye annet samtidig, og at det er store nivåforskjeller mellom elevene, som nevnt i innledningen.

Tidligere nevnte jeg at skrivekonstruktet har endret seg etter at Norskprøven ble digital. Pennington (2003, s. 287) hevder at litterasitetspraksisene er endret i dag, etter at datamaskinen har kommet inn overalt, og vi bruker den i alt arbeid med kommunikasjon og informasjon, både på jobb og privat. I den forbindelse er det mulig å hevde at elevene som ikke lærer å bruke digitale hjelpemidler på en adekvat og autentisk måte, ikke har fått en opplæring i funksjonelle skriveferdigheter, eller «…et ferdighetsnivå i norsk som setter dem i stand til å bruke eller bygge videre på sin kompetanse i utdanning, arbeid og samfunnsliv for øvrig» som det står i Læreplan i læreplan i norsk og samfunnskunnskap for voksne innvandrere (Vox, 2012, s. 7). Det er ikke urimelig å hevde at Vox ikke fullt ut har tatt deltakernes framtidige behov for digitale ferdigheter innover seg og implementert dem i læreplanen. Dermed fokuserer ikke skolene og lærerne nok på dette, og elevene blir kanskje ikke tilstrekkelig forberedt for verken Norskprøven eller livet etter prøven. Problemet med tidsbruk

som jeg nevnte over, blir naturligvis enda mer prekært, men skal man gi elevene det de har rett på og noe de har bruk for senere, er dette muligens nødvendig.

5.3 Anbefalinger for testing

Når Vox neste gang skal gjøre noen endringer på Norskprøven, bør de absolutt se på muligheter for rettferdig behandling av kandidater med svake PC-ferdigheter. Den siste setningen i formålsparagrafen for Det europeiske rammeverket for språk (CEFR) lyder slik: «Å avverge at folk som mangler de ferdighetene som kreves for å kommunisere i et interaktivt Europa, blir marginalisert» (Utdanningsdirektoratet, 2011, s. 4). Det krever ikke bare at Norskprøven blir endret, men også at læreplanen og informasjonen til skoler og lærere fullt ut implementerer de digitale kravene som mange av kandidatene sannsynligvis vil møte etter testen. Flere av deltakerne i mitt prosjekt hadde sannsynligvis ikke slike ferdigheter som CEFR har som formål i sitatet over ettersom de ikke ville ha nådd det nivået de ønsket i skriftlig framstilling på den testen jeg gjennomførte. Det er også sannsynlig at dette vil gjelde en del kandidater ellers i Norge hvert år.

Testing på PC er antagelig kommet for å bli, men det er store testorganisasjoner i andre land som tilbyr et alternativ til PC. For eksempel kan den verdensomspennende IELTS-testen[19] tas på papir hvis kandidaten ønsker det, nettopp av rettferdighetshensyn overfor kandidater med svake PC-ferdigheter (Davies, 2014). Også TOEFL-testen har vært på papir inntil nylig, og kan fortsatt tas på papir hvis internett mangler (ETS, 2016). Hvis konstruktet ikke er *skriving med PC* bør en tillate at noen skriver med penn på papir hvis de ønsker det, akkurat som at noen kandidater kan få oppgaver opplest hvis konstruktet ikke er å lese, men å skrive. Messick (1998, s. 16) kaller dette en *rettferdig tilpasning*, og sier at validiteten opprettholdes likevel. Fulcher (2014, s. 1559) hevder det samme. Den vanligste tilpasningen er *utvidet tid*, sier Fulcher. Kroll (1990, s. 140) hevder at utvidet tid kan gi kandidatene en sjanse til å forbedre seg når det er store krav til kognitiv kapasitet. Antagelig hadde mange av elevene som skrev på PC i mitt prosjekt fått bedre resultater med utvidet tid. Jeg nevnte tidligere den doble oversettelseshypotesen om andrespråksskrivere som skriver på PC (Wolfe & Manalo, 2004).

[19] IELTS er *International English Language Testing System*, drevet av Cambridge English Language Assessment for British Council. Disse konkurrerer med TOEFL fra ETS og Pearsons *Test of English Academic* om språktestmarkedet for engelsk som andrespråk (Spolsky, 2014, s. 1577).

Også Murphy & Yancey (2008) sier at testens validitet minker hvis tid er en avgjørende faktor for spesielle grupper, som nevnt.

Reglementet for Norskprøven (Vox, 2016e) opplyser om at det kan gis inntil 30 minutter ekstra tid etter søknad for «kandidater som har spesifikke lese- og skrivevansker, nedsatt syn, pollenallergi og andre typer helsemessige plager eller nedsatt funksjonsevne» (Vox, 2016e, s. 8). Det kan også tilrettelegges med «bruk av hjelpemidler i tradisjonell forstand» (ibid.), men i en fotnote påpekes det at ordbok eller stavekontroll ikke er tillatt. Hvilke hjelpemidler som derimot *er* tillatt, er dermed ganske uklart, og det gis ingen eksempler på dette. Å skrive med penn og papir er ikke nevnt som en tilrettelegging. Jeg mener at utvidet tid er en hensiktsmessig tilrettelegging i de tilfellene som er nevnt i reglementet fra Vox, men at svake ferdigheter i bruk av PC også burde gitt rett til utvidet tid. Det burde også vært mulig å skrive med penn og papir. Jeg tror at de eldste deltakerne med svake PC-ferdigheter som skrev på PC i mitt prosjekt, ville skrevet bedre tekster hvis de fikk utvidet tid eller kunne skrive med penn på papir. Det er ikke utenkelig at dette også kan gjelde en god del kandidater til Norskprøven hvert år.

Wolfe og Manalo (2005, s. 49) sier at kandidatene bør få velge skrivemodus når testen er «high-stake», men at man må være klar over at mange kandidater ikke kjenner sitt eget ferdighetsnivå på PC og at kandidater ofte tror at de får høyere skåre på PC-baserte tester. Med de affordansene som en PC tilbyr er det naturlig at noen kandidater kan føle at de skriver bedre på PC. Veiledning fra læreren og på Vox' nettside blir fortsatt viktig og bør utvides, også med veiledning på hvorvidt kandidaten bør velge PC eller penn og papir.

Jeg var tidligere inne på at Bachman og Palmer (1996) har autentisitet som en del av det de kaller testbrukbarhet. Autentisitet er også en del av konstruktvaliditeten, sier de (Bachman & Palmer, 1996, s. 42). En test er autentisk i den grad den er tilpasset virkelig språkbruk utenfor testen, fortsetter de. I en skole- eller jobbsituasjon har man tilgang til rettefunksjonen i Word hvis man jobber med tekstbehandling. Etter min mening vil det være mer autentisk å bruke rettefunksjonen også på Norskprøven. Norskprøven gjennomføres i et eget program, men det bør være mulig å implementere en rettefunksjon også i dette programmet. Som jeg viste i kapittel 2.4 gir bruk av PC klar forbedring i rettskriving og tegnsetting, gitt at eleven har fått instruksjon i dette. Og rettskriving og tegnsetting forbedres med rettefunksjonen (Pennington, 2003, s. 288). Antagelig ville mange av deltakerne i min undersøkelse, både yngre og eldre, fått bedre tekster hvis de fikk tilgang til rettefunksjonen. Selv om mange yrker i dag er basert

på at arbeidstakeren kan bruke en PC, er det fortsatt en del yrker som ikke krever det. Det er dessuten mulig å lære det man skal i jobben, det er tross alt ikke så mange arbeidsoppgaver som består i at arbeidstakeren skriver lengre tekster med tidsbegrensning alene og *uten noen hjelpemidler,* slik som det er i Norskprøven. Det vil derfor være mer autentisk å bruke rettefunksjonen på Norskprøven.

I det hele tatt bør kanskje selve skriveprøvens format og oppgavetyper endres, hvis man mener at konstruktet er endret. I rapporten fra Europarådet i forbindelse med overgangen til testing på PC i Europa, advarer René Meijer ved University of Derby mot å beholde oppgavene uforandret:

> Any instrument, in fact any implementation of an idea, is a compromise between the ideal and the constraints of reality. Assessments in this sense are no different. They are an attempt to translate a desired measurement into an instrument. When this instrument is replaced with another, for instance in the transition to computer-based assessment, we should be careful with simply implementing the old instrument into the new medium. Doing so would mean compromising the instrument with both the limitations of the old medium, and those of the new. (Meijer, 2009, s. 105).

Meijer mener altså at vi ikke bare kan bruke de samme oppgavene vi brukte i det gamle mediet, da vil vi få begrensningene *både* fra det gamle og det nye mediet. Vox har gjort noen endringer ved overgangen fra den gamle prøven til den nye prøven: De har laget en oppgave der man skal skrive til et bilde, men dette er en kjent oppgavetype som ikke gjør bruk av datamaskinens affordans. De har også endret utseendet på oppgave 2 og 3 slik at de grafisk likner et e-postprogram. Dette er ikke autentisk bruk av e-post, det er ikke noe annet enn det grafiske som skiller disse oppgaven og de gamle, der man skulle skrive et brev eller en beskjed. Selve oppgavetypene er altså gamle, det er for eksempel ingen bruk av interaktivitet eller hyperlenker, som jo er kvalitativt annerledes i digitale medier enn i papirbaserte. Dermed har Vox gjort nettopp det som Meijer (2009) advarer mot, de har beholdt gamle oppgaver i nytt medium. Å skrive på PC bør være noe kvalitativt annet enn å flytte oppgaver fra et medium til et annet. I et enda videre perspektiv er det dobbelt urettferdig mot elever som ikke har tilstrekkelige PC-ferdigheter og får svakere resultater på Norskprøven av den grunn: De får *verken* opplæring i funksjonelle skriveferdigheter *eller* ønsket nivå på skriftlig framstilling når læreplanen og prøven er som de er nå.

Kunnan (2005, s. 789) sier at man må ha hjelpe- eller støtteprosedyrer hvis det viser seg at konstruktvaliditeten ikke kan forsvares, og nevner *utvidet tid* eller et *annet format* som

eksempler på slike prosedyrer. Veilederen fra Europarådet til ansvarlige myndigheter i forbindelse med språkprøver ønsker at det skal finnes faste ordninger for «å sikre en rettferdig bedømmelse av personer med behov for spesiell tilrettelegging» (Balch et al., 2008). Den nevner en rekke tilretteleggingер som skal tilbys, men nevner ikke at det kan være en mulighet å skrive med penn og papir. Jeg tror dette er en svakhet.

Jeg nevnte i innledningen at det sannsynligvis er flere årsaker til at PC innføres i språktesting. Det er mer praktisk, som Moe (2003) og Bachman og Palmer (1996) viser at en språktest skal være. Innføringen av PC kan også skyldes såkalt *washback*, at elever og lærere «tvinges» til å bruke PC i undervisningen fordi det brukes PC på prøven (Carlsen, 2007; Shohamy, 2001). Slik washback er ikke nødvendigvis negativ, og når den kombineres med læreplanen er den kanskje også både fordelaktig og nødvendig. Likevel hevder flere at dette er å drive skoleutvikling «top-down» (Monsen, 2013, s. 40).

Som så ofte ellers, ligger det sannsynligvis også økonomiske motiver bak, særlig med økt innvandring som bakteppe (Scheuermann & Björnsson, 2009). Likevel må hensynet til økonomi og antall innvandrere balanseres mot rettferdig behandling av kandidatene, særlig de som ikke har tilstrekkelige ferdigheter på tekstbehandling, som jeg har vist i dette prosjektet. Slik balanse krever nesten alltid politiske bestemmelser, sier Messick (1998, s. 21). Det blir da forskernes jobb å informere politikere om alternativer og konsekvenser, fortsetter han. Og dette blir enda viktigere framover hvis norsk statsborgerskap blir avhengig av å bestå Norskprøven (Carlsen, 2011; Golden & Monsen, 2015). Slik er språktesting og forskning på språktesting i en særstilling: Mellom sårbare gruppers behov og fellesskapets ønsker om effektive og rettferdige språkprøver skal språktestingen komme med velfunderte metoder og relevant empiri. Shohamy (2009) og McNamara og Roever (2006, s. 213) minner om at språktester kan være kraftige politiske verktøy. Dette har ikke blitt mindre aktuelt og tydelig de siste årene. Sannsynligvis ønsker både politikere, forskere, lærere og elever å unngå at Norskprøven blir en digital sjibbolett som nekter noen elever adgang til rettigheter på bakgrunn av ukjente og urettferdige krav. Det er derfor all grunn til å minne om den dobbelte betydningen i Spolskys oppfordring fra 1981, gjentatt i 2014, når det gjelder språktesting: «Use with care».

Litteraturliste

Abrahamsson, N. (2009). *Andraspråksinlärning*. Lund: Studentlitteratur.

Abrahamsson, N., & Hyltenstam, K. (2013). Mognadsbegrensningar och den kritiska perioden för andraspråksinlärning. I K. Hyltenstam & I. Lindberg (Red.), *Svenska som andraspråk - i forskning, undervisning och samhälle* (2. utg., s. 221–258). Lund: Studentlitteratur.

Aliakbari, M. (2002). Writing in a foreign language: A writing problem or a language problem? *The Journal of PAAL*, *6*, 157–168.

ALTE. Materials for the guidance of test item writers (2005). Hentet fra http://www.alte.org/attachments/files/item_writer_guidelines.pdf

ALTE. (2007). ALTEs Praksiskodeks. [s.l]: Association of Language Testers in Europe. Hentet fra http://www.alte.org/attachments/files/code_practice_no.pdf

ALTE. Manual for language test development and examining (2011). Hentet fra http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

ALTE. (2016). ALTE Membership. Hentet 22. oktober 2016, fra http://www.alte.org/membership

Arthur, B. (1979). Short-term changes in EFL composition skills. I C. A. Yorio, K. Perkins, & J. Schachter (Red.), *On TESOL '79: The Learner in Focus* (s. 330–342). Washington, D.C.: TESOL.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*(4), 671–704.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Balch, A., Corrigan, M., Gysen, S., Kuijper, H., Perlmann-Balme, M., Roppe, S., … Zeidler, B. (2008). *Språkprøver i forbindelse med sosial integrering og statsborgerskap - en veiledning til ansvarlige myndigheter*. [s.l]: Europarådet.

Bangert-Drowns, R. L. (1993). The Word Processor as an Instructional Tool: A Meta-Analysis of Word Processing in Writing Instruction. *Review of Educational Research*, *63*(1), 69–93.

Bardovi-Harlig, K. (1992). A Second Look at T-unit Analysis: Reconsidering the Sentence. *TESOL Quarterly*, *26*(2), 390–395.

Baron, N. S. (2008). *Always On: Language in an Online and Mobile World*. New York: Oxford University Press.

Berge, A. L. (1999). Norsk som andrespråk i videregående skole. I J. E. Hagen & K. Tenfjord (Red.), *Andrespråksundervisning. Teori og praksis* (s. 125–149). Oslo: Gyldendal Akademisk.

Berge, K. L., & Tønnesson, J. L. (2007). Skriveopplæring og skriveforskning for nåtid og framtid. Hvor går veien? I S. Matre & T. Løkensgard Hoel (Red.), *Skrive for nåtid og framtid 1. Skriving i arbeidsliv og skole* (s. 29–36). Trondheim: Tapir Akademiske.

Berggreen, H., & Tenfjord, K. (2007). *Andrespråkslæring* (2. utg.). Oslo: Gyldendal Akademisk.

Biber, D., & Conrad, S. (2009). *Register, genre and style*. Cambridge: Cambridge University Press.

Biber, D., Gray, B., & Poonpon, K. (2011). Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? *TESOL Quarterly*, *45*(1), 5–35.

Bisaillon, J. (1999). Effects of the teaching of revision strategies in a computer-based environment. I M. C. Pennington (Red.), *Writing in an Electronic Medium: Research with Language Learners* (s. 131–157). Houston, TX: Athelstan.

Bjørkeng, B. (2013). *Ferdigheter i voksenbefolkningen. Resultater fra den internasjonale undersøkelsen om lese- og tallforståelse (PIAAC)*. Oslo/Kongsvinger: SSB.

Björnsson, C. H. (1968). *Läsbarhet*. Stockholm: Liber.

Broeder, P., Extra, G., & van Hout, R. (1993). Richness and variety in the developing lexicon. I C. Perdue (Red.), *Adult language acquisition: cross-linguistic perspectives. Volume 1: Field methods.* (s. 145–163). Cambridge: Cambridge University Press.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Carlsen, C. (2000). Den objektive språktesten - mål eller minne? I R. B. Brodersen & T. Kinn (Red.), *Språkvitskap og vitskapsteori. Ti nye vitskapsteoretiske innlegg*. (s. 69–88). Larvik: Ariadne forlag.

Carlsen, C. (2003). Et forsøk på å forklare sensorenighet. I W. Vagle (Red.), *Vurdering av språkferdighet, rapport nr. 1 fra KAL* (s. 107–122). Trondheim: Institutt for språk- og kommunikasjonsstudier, NTNU.

Carlsen, C. (2005). Karakterens gåtefulle karakter. *Nordica Bergensia*, *32*, 5–16.

Carlsen, C. (2007). Language testing - a matter of ethics. I C. Carlsen & E. Moe (Red.), *A Human Touch to Language Testing. A collection of essays in honour of Reidun Oanæs Andersen on the occasion of her retirement June 2007* (s. 97–107). Oslo: Novus.

Carlsen, C. (2011). Språkprøver - redskap for integrering? I *Godt no(rs)k? - om språk og integrering* (s. 97–102). Oslo: IMDI.

Carlsen, C. (2012). Rammeverket, referansenivåbeskrivelser og innlærerkorpuset ASK. I C. Carlsen (Red.), *Norsk profil. Det felles europeiske rammeverket spesifisert for norsk. Et første steg.* (s. 15–48). Oslo: Novus.

Carlsen, C., & Moe, E. (2014). Assessing Norwegian. I A. J. Kunnan (Red.), *The companion to language assessment, vol. 4: Assessment around the world* (s. 2031–2037). Malden, MA: Wiley Blackwell.

Carroll, J. B. (1967). On sampling from a lognormal model of word frequency distribution. I H. Kucera & W. N. Francis (Red.), *Computational analysis of present-day American English* (s. 406–424). Providence, RI: Brown University Press.

Carson, J., & Kuehn, P. (1992). Evidence of transfer and loss in developing second language writers. *Language Learning*, *42*(2), 157–182.

Casanave, C. P. (1994). Language Development in Students' Journals. *Journal of Second Language Writing*, *3*(3), 179–201.

Chapelle, C. A. (2012). Conceptions of validity. I G. Fulcher & F. Davidson (Red.), *Routledge handbook of language testing* (s. 21–33). Abingdon: Routledge.

Chapelle, C. A., & Douglas, D. (2006). *Assessing Language Trough Computer Technology*. Cambridge: Cambridge University Press.

Clark, R. E. (1985). Evidence for confounding in computer-based instruction studies: Analyzing the meta-analyses. *Educational Communication and Technology Journal*, *33*(4), 249–262.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. utg.). Newbury Park, CA: SAGE.

Collins, P., Hwang, J. K., Zheng, B., & Warschauer, M. (2013). Writing with Laptops: A Quasi-Experimental Study. *Writing & Pedagogy*, *5*(2), 203–230.

Collot, M., & Belmore, N. (1996). Electronic language: A new variety of English. I S. C. Herring (Red.), *Computer-Mediated Communication* (s. 13–28). Amsterdam: John Benjamins.

Creaza. (2016). Creaza Cartoonist.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*, 119–135.

Cumming, A. (1996). Introduction: The concept of validation in language testing. I A. Cumming & R. Berwick (Red.), *Validation in language testing* (s. 1–14). Clevedon: Multilingual Matters.

Cumming, A. (1997). The testing of writing in a second language. I C. Clapham & D. Corson (Red.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (s. 51–63). Dordrecht: Kluwer Academic Publishers.

Cummins, J. (1981). Age on Arrival and Immigrant Second Language Learning in Canada: A Reassessment. *Applied Linguistics*, *11*(2), 132–149.

Cummins, J. (2008). BICS and CALP: Empirical and Theoretical Status of the Distinction. I B. Street & N. H. Hornberger (Red.), *Encyclopedia of language and education, vol. 2: Literacy* (2. utg., s. 71–83). New York: Springer.

Danbolt, A. M. V. (2004). Repetisjon versus ekspansjon. En studie av ordforrådet i to tekster skrevet av minoritetselever på ungdomstrinnet. *NOA norsk som andrespråk*, *25*, 1–16.

Davies, A. (2014). Fifty years of language assessment. I A. J. Kunnan (Red.), *The companion to language assessment, vol. 1: Abilities, contexts and learners* (s. 1–19). Malden, MA: Wiley Blackwell.

Dewaele, J.-M., & Pavlenko, A. (2003). Productivity and Lexical Diversity on Native and Non-Native Speech: A Study of Cross-cultural Effects. I V. J. Cook (Red.), *Second Language Acquisition, 3: Effects of the Second Language on the First* (s. 120–141). Clevedon: Multilingual Matters.

Difi. (2016). Klart språk. Hentet 13. oktober 2016, fra https://www.difi.no/fagomrader-og-tjenester/klart-sprak-og-brukerinvolvering/klart-sprak

Dischler, R. (2011). Norske arbeidsgivere stiller strenge språkkrav. I *Godt no(rs)k? - om språk og integrering* (s. 66–71). Oslo: IMDI.

Dunn, B., & Reay, D. (1989). Word processing and the keyboard: Comparative effects of transcription on achievement. *The Journal of Educational Research*, *82*(4), 237–245.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220–242.

Dyck, J. L., & Al-Awar Smither, J. (1996). Older Adults' Acquisition of Word Processing: The Contribution of Cognitive Abilities and Computer Anxiety. *Computers in Human Behavior*, *12*(1), 107–119.

Dysthe, O., & Hertzberg, F. (2007). Kunnskap om skriving i utdanning og yrkesliv - hvor står vi i dag? I S. Matre & T. Løkensgard Hoel (Red.), *Skrive for nåtid og framtid 1. Skriving i arbeidsliv og skole* (s. 10–28). Trondheim: Tapir Akademiske.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Egghe, L. (2007). Untangling Herdan's Law and Heaps' Law: Mathematical and Informetric Arguments. *Journal of the American Society for Information Science and Technology*, *58*(5), 702–709.

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL

compositions. *Journal of Second Language Writing*, *4*(2), 139–155.
Eriksson, T., & Carlsen, C. (2012). Rammeverket, feilmengde og feilmønster. I C. Carlsen (Red.), *Norsk profil. Det felles europeiske rammeverket spesifisert for norsk. Et første steg.* (s. 245–267). Oslo: Novus.
ETS. (2014). ETS Standards for Quality and Fairness. [s.l]: Educational Testing Service.
ETS. (2016). Frequently Asked Questions about the TOEFL iBT® Test. Hentet 24. oktober 2016, fra https://www.ets.org/toefl/ibt/faq/
Faarlund, J. T., Lie, S., & Vannebo, K. I. (1997). *Norsk referansegrammatikk.* Oslo: Universitetsforlaget.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*, 175–191.
Field, A. (2000). *Discovering Statistics using SPSS for Windows*. London: SAGE.
Fife-Schaw. (2016). Statistics FAQ. Hentet 10. september 2016, fra http://www.surrey.ac.uk/psychology/current/statistics/index.htm
Fjørtoft, H. (2014). *Norskdidaktikk*. Bergen: LNU/Fagbokforlaget.
Flower, J. R., & Hayes, L. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.
Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375.
Friedlander, A. (1990). Composing in English: effects of a first language on writing in English as a second language. I B. Kroll (Red.), *Second language writing: Research insights for the classroom* (s. 109–125). Cambridge: Cambridge University Press.
Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18.
Fulcher, G. (2000). The «communicative» legacy in language testing. *System*, *28*, 483–497.
Fulcher, G. (2014). Language Testing in the Dock. I A. J. Kunnan (Red.), *The companion to language assessment, vol. 3: Evaluation, methodology, and interdisciplinary themes* (s. 1553–1570). Malden, MA: Wiley Blackwell.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Abingdon: Routledge.
Gabrielsen, E., & Lagerstrøm, B. O. (2007). *Med annen bakgrunn. Lese- og regneferdigheter blant norske innvandrere*. Stavanger: Lesesenteret UiS.
Gaetano, J. (2013). Holm-Bonferroni sequential correction: An EXCEL calculator (1.1). Hentet 12. august 2016, fra http://tinyurl.com/hdp7fff
Goldberg, A., Russell, M., & Cook, A. (2003). The Effect of Computers on Student Writing: A Meta-Analysis of Studies from 1992 to 2002. *The Journal of Technology, Learning, and Assessment*, *2*(1), 2–51.
Golden, A. (2014). *Ordforråd, ordbruk og ordlæring* (4. utg.). Oslo: Gyldendal Akademisk.
Golden, A., Hvenekilde, A., & Ryen, E. (1995). Forord. *NOA norsk som andrespråk*, *18*.
Golden, A., & Hvistendahl, R. (2015). Forskning på andrespråksskriving i Skandinavia, med vekt på de norske studiene. I A. Golden & E. Selj (Red.), *Skriving på norsk som andrespråk: Vurdering, opplæring og elevenes stemmer* (s. 231–246). Oslo: Cappelen Damm Akademisk.
Golden, A., Kulbrandstad, L. I., & Tenfjord, K. (2007). Norsk andrespråksforskning - utviklingslinjer fra 1980 til 2005. *Nordand*, *2*(1).
Golden, A., & Monsen, M. (2015). Vurdering av tekster skrevet til norskprøvene for voksne. I A. Golden & E. Selj (Red.), *Skriving på norsk som andrespråk: Vurdering, opplæring og elevenes stemmer* (s. 201–215). Oslo: Cappelen Damm Akademisk.

Grabe, W., & Kaplan, R. B. (1996). *Theory and Practice of Writing*. Harlow: Longman.

GraphPad. (2016). Interpreting results: Kruskal-Wallis test. Hentet 14. september 2016, fra http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_kruskal-wallis_test_works.htm

Hagen, J. E. (2005). Refleksjoner gjennom andrespråksprismet. I S. Lie, G. Nedrelid, & H. Omdal (Red.), *MONS 10. Utvalde artiklar frå det tiande Møte om norsk språk i Kristiansand 2003.* Kristiansand: Høyskoleforlaget.

Halliday, M. A. K. (1987). Spoken and written modes of meaning. I R. Horowitz & S. J. Samuels (Red.), *Comprehending oral and written language* (s. 55–82). New York: Academic Press.

Halliday, M. A. K. (1989). *Spoken and Written Language* (2.). Oxford: Oxford University Press.

Halvorsen, B. (2003). Vurdering av muntlig språkferdighet. I W. Vagle (Red.), *Vurdering av språkferdighet, rapport nr. 1 fra KAL* (s. 135–141). Trondheim: Institutt for språk- og kommunikasjonsstudier, NTNU.

Hammarberg, B. (2013). Teoretiske ramar för andraspråksforskning. I K. Hyltenstam & I. Lindberg (Red.), *Svenska som andraspråk - i forskning, undervisning och samhälle* (2. utg., s. 27–83). Lund: Studentlitteratur.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. I B. Kroll (Red.), *Second language writing: Research insights for the classroom* (s. 69–87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). The Writer's Knowledge and Our Knowledge of the Writer. I L. Hamp-Lyons (Red.), *Assessing Second Language Writing in Academic Contexts* (s. 51–68). Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. I B. Kroll (Red.), *Exploring the dynamics of second language writing* (s. 162–189). New York: Cambridge University Press.

Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL*, *6*(1), 52–72.

Hatch, E., & Lazaraton, A. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. New York: Newbury House.

Henning, G. (1991). Issues in Evaluating and Maintaining an ESL Writing Assessment Program. I L. Hamp-Lyons (Red.), *Assessing Second Language Writing in Academic Contexts* (s. 279–291). Norwood, NJ: Ablex Publishing Corporation.

Herdan, G. (1964). *Quantitative Linguistics*. London: Butterworth.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, *37*(2), 275–301.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandianavian Journal of Statistics*, *6*(2), 65–70.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2), 1–50.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. I A. Housen, F. Kuiken, & I. Vedder (Red.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. LL/LT 32.* (s. 1–20). Amsterdam: John Benjamins.

Hudson, R. (2009). Measuring maturity. I R. Beard, D. Myhill, J. Riley, & M. Nystrand (Red.), *The Sage Handbook of Writing Development* (s. 349–362). London: SAGE.

Hultman, T. G., & Westman, M. (1977). *Gymnasistsvenska*. Lund: Liber Läromedel.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels. NCTE Research*

Report no. 3. Champaign, IL, USA: National Council of Teachers of English.
Huot, B., & O'Neill, P. (2009). An introduction to writing assessment theory and practice. I B. Huot & P. O'Neill (Red.), *Assessing writing: A critical sourcebook* (s. 1–9). Boston, MA: Bedford/St. Martins.
Hyland, K. (2003). *Second language writing*. Cambridge: Cambridge University Press.
Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of multilingual and multicultural development*, *9*(1/2), 67–84.
Hymes, D. (1972). On communicative competence. I J. B. Pride & J. Holmes (Red.), *Sociolinguistics: Selected readings* (s. 269–293). Harmondsworth: Penguin.
Hård af Segerstad, Y., & Sofkova Hashemi, S. (2006). Learning to write in the information age: A case study of schoolchildren's writing in Sweden. I L. van Waes, M. Leijten, & C. M. Neuwirth (Red.), *Writing and digital media*. Oxford: Elsevier Ltd.
IBM. (2015). IBM SPSS 23. IBM.
IMDI. (2011). Finansiering av norskopplæring. I *Godt no(rs)k? - om språk og integrering* (s. 40–41). Oslo: Integrerings- og mangfoldsdirektoratet.
Intaraprawat, P., & Steffensen, M. S. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing*, *4*(3), 253–272.
internetlivestats.com. (2016). Internet Users by Country (2016). Hentet 26. juli 2016, fra http://www.internetlivestats.com/internet-users-by-country/
Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, *4*(1), 51–69.
Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84.
Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*(Suppl. 1), 87–106.
Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377–403.
Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. I *Working Papers* (Bd. 53, s. 61–79). Lund University.
Jølbo, I. D. (2016). *Identitet, stemme og aktørskap i andrespråksskriving : en undersøkelse av skriving som meningsskaping blant elever med somalisk bakgrunn i norskfaget i grunnskoleopplæringen for minoritetsspråklig ungdom*. Doktorgradsavhandling, ILN, Universitetet i Oslo.
Kaplan, R. B. (1966). Cultural Thought Patterns in Inter-Cultural Education. *Language Learning*, *16*(1–2), 1–20.
Kellogg, R. T. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, *114*(2), 175–191.
Kokkinakis, S. J., & Magnusson, U. (2011). Computer based quantitative methods applied to first and second language writing. I *Young Urban Swedish: Variation and change in multilingual settings* (s. 105–124). Göteborg: Universitetet i Göteborg.
Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*, 148–161.
Kristiansen, J. E. (2010). Mot normalt: om gjennomsnittet. Hentet 6. september 2016, fra https://www.ssb.no/sosiale-forhold-og-kriminalitet/artikler-og-publikasjoner/mot-normalt-om-gjennomsnittet
Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. I B. Kroll (Red.), *Second language writing: Research insights for the classroom* (s. 141–154). Cambridge: Cambridge University Press.
Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. I A.

Housen, F. Kuiken, & I. Vedder (Red.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. LL/LT 32.* (s. 143–170). Amsterdam: John Benjamins.

Kulbrandstad, L. A. (2005). *Språkets mønstre* (3. utg.). Oslo: Universitetsforlaget.

Kunnan, A. J. (2005). Language Assessment From a Wider Context. I E. Hinkel (Red.), *Handbook of research in second language teaching and learning* (s. 779–794). Mahwah, NJ: Erlbaum.

Lado, R. (1961). *Language Testing: the Construction and Use of Foreign Language Tests*. New York: McGraw-Hill.

Larsen-Freeman, D. (1978). An ESL Index of Development. *Tesol Quarterly*, *12*(4), 439–448.

Larsen-Freeman, D. (1983). Assessing Global Second Language Proficiency. I *Classroom Oriented Research in Second Language Acquisition* (s. 287–304). Rowley, MA: Newbury House.

Larsen-Freeman, D. (1997). Chaos/Complexity Science and Second Language Acquisition. *Applied Linguistics*, *18*(2), 141–165.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579–589.

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

Larson-Hall, J. (2011). How to Run Statistical Analyses. I A. Mackey & S. M. Gass (Red.), *Research Methods in Second Language Acquisition: A Practical Guide* (s. 245–274). Malden, MA: Wiley Blackwell.

Larsson, K. (1984). *Skrivförmåga. Studier i svenskt elevspråk.* Malmö: Liber.

Laufer, B. (1994). The Lexical Profile of Second Language Writing: Does It Change Over Time? *RELC Journal*, *25*(2), 21–33.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255–271.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, *31*(4), 537–561.

Lennon, P. (1991). Error: Some Problems of Definition, Identification, and Distinction. *Applied Linguistics*, *12*(2), 180–196.

Levene, H. (1960). Robust tests for equality of variances. I I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Red.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (s. 278–292). Stanford, CA: Stanford University Press.

Lie, S., Kjærnsli, M., Roe, A., & Turmo, A. (2001). Godt rustet for framtida? Norske 15-åringers kompetanse i lesing og realfag i et internasjonalt perspektiv. *Acta Didacta*, *4*.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund: Gleerup.

Lintunen, P., & Mäkilä, M. (2014). Measuring Syntactic Complexity in Spoken and Written Learner Language: Comparing the Incomparable? *Research in Language*, *12*(4), 377–399.

Long, M. H. (2003). Stabilization and fossilization in interlanguage. I C. J. Doughty & M. H. Long (Red.), *The handbook of second language acquisition* (s. 487–535). Malden, MA: Blackwell.

Lowie, W., & Seton, B. (2013). *Essential Statistics for Applied Linguistics*. Basingstoke: Palgrave Macmillan.

Lundberg, I., & Reichenberg, M. (2009). *Vad är lättläst?* Härnösand: Specialpedagogiska

skolmyndigheten.
Macaro, E. (2003). *Teaching and learning a second language: A review of recent research*. London and New York: Continuum.
MacArthur, C. A. (2006). The effects of new technologies on writing and writing processes. I C. A. MacArthur, S. Graham, & J. Fitzgerald (Red.), *Handbook of writing research* (s. 248–262). New York: The Guilford Press.
Mackey, A., & Gass, S. M. (2005). *Secong Language Research: Methodology and Design*. Mahwah, NJ: Lawrence Erlbaum.
MacWhinney, B. (2003). CHILDES - Child Language Data Exchange System. Hentet 8. oktober 2016, fra http://childes.psy.cmu.edu/
Madsen, H. S. (1991). Computer-Adaptive Testing of Listening and Reading Comprehension: The Brigham Young University Approach. I P. Dunkel (Red.), *Computer-Assisted Language Learning and Testing: Research Issues and Practice* (s. 237–257). New York: Newbury House.
Magnusson, U. (2013). Skrivande på ett andraspråk. I K. Hyltenstam & I. Lindberg (Red.), *Svenska som andraspråk - i forskning, undervisning och samhälle* (2., s. 633–660). Lund: Studentlitteratur.
Martens, G. (2009). *Hva kan ordforrådet i skriftlige tekster til andrespråkselever fortelle oss om deres læreforutsetninger? Forskningsrapport fra Universitetet i Tromsø*. Tromsø.
Matsuda, P. K. (1997). Contrastive Rhetoric in Context: A Dynamic Model of L2 Writing. *Journal of Second Language Writing*, *6*(1), 45–60.
Matsuda, P. K. (2003). Second language writing in the twentieth century: A situated historical perspective. I B. Kroll (Red.), *Exploring the dynamics of second language writing* (s. 15–34). New York: Cambridge University Press.
McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42*(2), 381–92.
McKee, G., Malvern, D., & Richards, B. (2000). Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Linguistic Computing*, *15*(3), 323–337.
McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.
McNamara, T. (2003). Looking back, looking forward: rethinking Bachman. *Language Testing*, *20*(4), 466–473.
McNamara, T. (2005). 21st Century Shibboleth: Language Tests, Identity and Intergroup Conflict. *Language Policy*, *4*, 351–370.
McNamara, T., & Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA: Blackwell.
Medienorge. (2016). PC-bruk en gjennomsnittsuke. Hentet 26. juli 2016, fra http://www.medienorge.uib.no/statistikk/medium/ikt/250
Meijer, R. (2009). Transititon to Computer-based Assessment: Motivations and considerations. I F. Scheuermann & J. Björnsson (Red.), *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing* (s. 104–107). Luxembourg: Europarådet.
Messick, S. (1989). Validity. I R. L. Linn (Red.), *Educational Measurement, Third Edition* (s. 13–103). New York: ACE/Macmillan.
Messick, S. (1998). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. ETS research report, 48.* Princeton, NJ.
Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations,

and textbooks across Europe. I I. Bartning, M. Martin, & I. Vedder (Red.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research, Eurosla monographs 1* (s. 211–231). [s.l]: Eurosla.

Moe, E. (2002). Syntaktiske trekk i norsk som andrespråk på eit mellomnivå. *Nordica Bergensia, 26*, 183–209.

Moe, E. (2003). Den gode prøven - finst han? I W. Vagle (Red.), *Vurdering av språkferdighet, rapport nr. 1 fra KAL* (s. 123–134). Trondheim: Institutt for språk- og kommunikasjonsstudier, NTNU.

Moe, E. (2008). Språkprøvar, språkkompetanse og Det europeiske rammeverket. I C. Carlsen, E. Moe, R. Oanæs Andersen, & K. Tenfjord (Red.), *Banebryter og brobygger i andrespråksfeltet: En samling artikler i anledning Jon Erik Hagens 60-årsdag* (s. 134–145). Oslo: Novus.

Monsen, M. (2008). *Kommunikativ funksjonalitet kontra formell korrekthet. Ordforrådsrikdom og leksikalske og formelle feil i besvarelser til Norskprøve 3 for voksne innvandrere*. Masteravhandling, Høgskolen i Hedmark.

Monsen, M. (2013). *Store forventninger? Læreroppfatninger om eksterne leseprøver*. Doktorgradsavhandling, Universitetet i Oslo.

Monsen, M. (2015). Andrespråksdidaktisk forskning på voksenopplæring i Norge: En oversikt fra 1985 til i dag. *NOA norsk som andrespråk, 30*(1–2), 373–392.

Murphy, S., & Yancey, K. B. (2008). Construct and Consequence: Validity in Writing Assessment. I C. Bazerman (Red.), *Handbook of research on writing: History, society, school, individual, text* (s. 365–385). New York: Lawrence Erlbaum.

Neu, J., & Scarcella, R. (1991). Word Processing in the ESL Writing Classroom: A Survey of Student Attitudes. I P. Dunkel (Red.), *Computer-Assisted Language Learning and Testing: Research Issues and Practice* (s. 169–187). New York: Newbury House.

Nordstokke, D. W., Zumbo, B. D., Cairns, S. L., & Saklofske, D. H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research & Evaluation*, *16*(5).

Norris, J. M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics*, *30*(4), 555–578.

Norsk Språktest. (2016). Hentet fra http://www.uib.no/lle/30641/norsk-spr%C3%A5ktest

Nyman, S. (2014). *Utvärdering av kustväderuppläsning: Vilken språkprofil förstår användaren bäst?* Linköpings universitet.

Nystrand, M., Greene, S., & Wiemelt, J. (1993). Where did composition studies come from? An intellectual history. *Written Communication*, *10*(3), 267–333.

Nyström, C. (2000). *Gymnasisters skrivande: En studie av genre, textstruktur och sammanhang*. Uppsala universitet.

Ongstad, S. (2002). Positioning Early Research on Writing in Norway. *Written Communication*, *19*(3), 345–381.

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, *24*(4), 492–518.

Ortega, L. (2009). Studying Writing Across EFL Contexts: Looking Back and Moving Forward. I R. M. Manchón (Red.), *Writing in Foreign Language Contexts: Learning, Teaching, and Research*. Bristol: Multilingual Matters.

Pallotti, G. (2009). CAF: Defining, Refining, and Differentiating Constructs. *Applied Linguistics*, *30*(4), 590–601.

Pedersen, J. (2008). Konstruktvalidering av språkprøver - forholdet mellom spørsmål og svar. I *Banebryter og brobygger i andrespråksfeltet: En samling artikler i anledning Jon Erik Hagens 60-årsdag* (s. 88–100). Oslo: Novus.

Pennington, M. C. (2003). The impact of the computer in second language writing. I B. Kroll

(Red.), *Exploring the dynamics of second language writing* (s. 287–310). Cambridge: Cambridge University Press.

Plakans, L. (2014). Written Discourse. I A. J. Kunnan (Red.), *The companion to language assessment, vol. 3: Evaluation, methodology, and interdisciplinary themes* (s. 1390–1402). Malden, MA: Wiley Blackwell.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, *47*(1), 101–143.

Polio, C. (2003). Research on second language writing: An overview of what we investigate and how. I B. Kroll (Red.), *Exploring the dynamics of second language writing* (s. 35–65). New York: Cambridge University Press.

Polio, C. (2012). The acquisition of second language writing. I S. M. Gass & A. Mackey (Red.), *The Routledge handbook of second language acquisition* (s. 319–334). Abingdon: Routledge.

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27.

Reynolds, D. W. (2005). Linguistic correlates of second language literacy development: Evidence from middle-grade learner essays. *Journal of Second Language Writing*, *14*, 19–45.

Roblyer, M. D., Castine, W. H., & King, F. J. (1988). *Assessing the impact of computer-based instruction*. New York: The Haworth Press.

Rowe Krapels, A. (1990). An overview of second language writing process research. I B. Kroll (Red.), *Second language writing: Research insights for the classroom* (s. 37–56). Cambridge: Cambridge University Press.

Rusmin, R. S. (1999). Patterns of adaptation to a new writing environment: The experience of word processing by mature second language writers. I M. C. Pennington (Red.), *Writing in an Electronic Medium: Research with Language Learners* (s. 183–227). Houston, TX: Athelstan.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, *17*(4), 688–690.

Saeed, J. I. (2003). *Semantics* (2. utg.). Malden, MA: Blackwell.

Sandvik, M. (2008). Digitale verktøy i det flerkulturelle klasserommet. I E. Selj & E. Ryen (Red.), *Med språklige minoriteter i klassen* (2. utg., s. 157–175). Oslo: Cappelen Akademisk.

Scheuermann, F., & Björnsson, J. (2009). *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*. Luxembourg: Europarådet.

Seimyr, G. Ö. (udatert). LIX räknare. Hentet 25. mars 2016, fra http://lix.se

Seliger, H. W., & Shohamy, E. (1989). *Second Language Research Methods*. Oxford: Oxford University Press.

Selj, E. (2008). Skriving når norsk er andrespråk. I E. Selj & E. Ryen (Red.), *Med språklige minoriteter i klassen* (2. utg., s. 131–156). Oslo: Cappelen Akademisk.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. London and New York: Routledge.

Shohamy, E. (2009). Language Tests for Citizenship, Immigration, and Asylum. *Language Assessment Quarterly*, *6*, 1–5.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, *27*(4), 657–677.

Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance: A synthesis of the Ealing research. I *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. LL/LT 32.* (s. 199–220).

Amsterdam: John Benjamins.
Spolsky, B. (1978). *Approaches to language testing: Advances in language testing series: 2*. Arlington, VA: Center for applied linguistics.
Spolsky, B. (2008). Language assessment in historical and future perspective. I E. Shohamy & N. H. Hornberger (Red.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (2. utg., s. 445–454). New York: Springer.
Spolsky, B. (2014). The Influence of Ethics in Language Assessment. I A. J. Kunnan (Red.), *The companion to language assessment, vol. 3: Evaluation, methodology, and interdisciplinary themes* (s. 1571–1585). Malden, MA: Wiley Blackwell.
SSB. (2016). Norskopplæring for voksne innvandrere. Hentet 2. november 2016, fra http://tinyurl.com/zfsqv32
Søyland, A., & Fretland, J. O. (2015). *Norske skriveregler. Reglene du trenger for å skrive på papir og skjerm*. [s.l]: Samlaget.
Tenfjord, K., Hagen, J. E., & Johansen, H. (2009). Norsk andrespråkskorpus (ASK) - design og metodiske forutsetninger. *NOA norsk som andrespråk, 25*(1), 52–81.
Totland, R. K., & Lauvik, H. (2012). Ordforråd og ordvalg fra A2 til C1. I C. Carlsen (Red.), *Norsk profil. Det felles europeiske rammeverket spesifisert for norsk. Et første steg.* (s. 189–221). Oslo: Novus.
Ure, J. (1971). Lexical density and register differentiation. I G. Perren & J. L. M. Trim (Red.), *Applications of linguistics* (s. 443–452). London: Cambridge University Press.
Utdanningsdirektoratet. (2011). Det felles europeiske rammeverket for språk. Læring, undervisning, vurdering. Hentet fra http://www.udir.no/Upload/Verktoy/5/UDIR_Rammeverk_sept_2011_web.pdf
Vagle, W. (2005). Studie 10: Tekstlengde + ordlengdesnitt = kvalitet? Hva kvantitative kriterier forteller om avgangselevenes skriveprestasjoner. I K. L. Berge, L. S. Evensen, F. Hertzberg, & W. Vagle (Red.), *Ungdommers skrivekompetanse. Bind II: Norskeksamen som tekst* (s. 303–386). Oslo: Universitetsforlaget.
van Dijk, J. A. G. M. (2006). Digital divide research, achievements and shortcomings. *Poetics, 34*, 221–335.
van Waes, L. (1994). Computers and writing. Implications for the teaching of writing. I K.-H. Pogner (Red.), *More about writing* (s. 41–61). Odense: Institute of Language and Communication, Odense University.
Veazie, P. J. (2006). When to Combine Hypotheses and Adjust for Multiple Tests. *HSR: Health Services Research, 41*(3), 804–818.
Viberg, Å. (2004). Lexikal utveckling i ett andraspråk. I K. Hyltenstam & I. Lindberg (Red.), *Svenska som andraspråk - i forskning, undervisning och samhälle* (s. 197–220). Lund: Studentlitteratur.
Vollan, S. (2014). De nye prøvene. Erfaringer så langt og videre planer. Hentet 29. oktober 2016, fra https://www.fylkesmannen.no/Documents/Dokument FMBU/Kursdokumenter/2014/VOX - De nye pr%C3%B8vene %E2%80%93 erfaringer og planer.pdf
Vox. (2011). Norskopplæring for innvandrere. I *Godt no(rs)k? - om språk og integrering* (s. 12–17). Oslo: IMDI.
Vox. (2012). Læreplan i norsk og samfunnskunnskap for voksne innvandrere. Oslo: Vox.
Vox. (2016a). Eksempelprøve. Hentet fra https://test.flexiteexam.com/adapt-it/Begin#/assessment/be98a105562870ca0156698655001795
Vox. (2016b). Nivåvelger for norsk. Hentet 12. oktober 2016, fra https://adaptit.enovate.no/latest/Norskniva
Vox. (2016c). Norskprøven - innhold. Hentet fra http://www.vox.no/Norsk-og-samfunnskunnskap/Norskprove/#ob=9067

Vox. (2016d). Norskprøveresultater. Hentet fra http://tinyurl.com/jn87ooa
Vox. (2016e). Reglement for gjennomføring av norskprøven for voksne innvandrere. Hentet fra http://tinyurl.com/hggx37w
Vox. (2016f). Vox statistikkbanken - forklaring. Hentet 31. oktober 2016, fra http://tinyurl.com/h46o29t
Vox. (2016g). Vurderingsskjema for norskprøven, delprøve i skriftlig framstilling. Hentet 13. november 2016, fra http://tinyurl.com/haol4a5
Wang, H., & Shin, C. D. (2009). Computer-Based and Paper-Pencil Test Comparability Studies. *Pearson Test, Measurement, and Research Services Bulletin*, (9), 1–6.
Warschauer, M., & Liaw, M.-L. (2010). *Emerging Technologies in Adult Literacy and Language Education*. Washington, D.C.
Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.
Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language learning & technology*, *8*(1), 53–65.
Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL writing scores*. Princeton, NJ.
Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i.
Zimmerman, D. W. (2004). Inflation of Type I Error Rates by Unequal Variances Associated with Parametric, Nonparametric, and Rank-Transformation Tests. *Psicológica*, *25*, 103–133.
Östlund-Stjärnegårdh, E. (2002). *Godkänd i svenska? Bedömning och analys av gymnasieelevers texter*. Doktorgradsavhandling, Uppsala universitet.

Vedlegg 1: Prøve i skriftlig framstilling i dette prosjektet

Skriftlig A2/B1-prøve

Delprøven i skriftlig framstilling A2/B1 har 3 oppgaver. **Du må svare på alle oppgavene for å få en vurdering.**

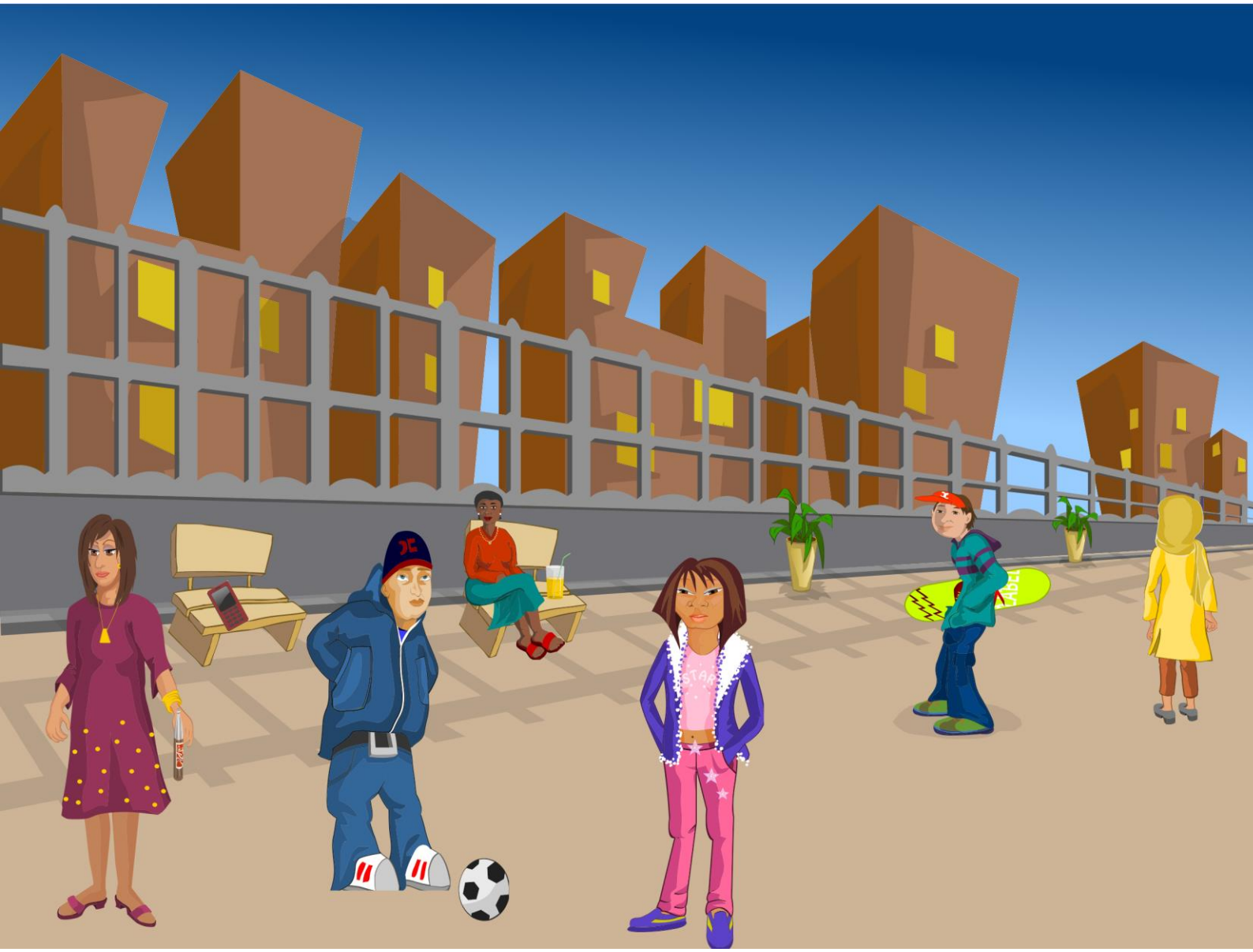
Spør prøveleder hvis du ikke forstår oppgavene.

Oppgave 1 – beskrive et bilde (80-100 ord)

Oppgave 2 – fortelle om et kjent tema (80-200 ord)

Oppgave 3 – uttrykke egne meninger (for B1 bør du skrive minst 80 ord på denne oppgaven)

Du får 90 minutter til å svare på oppgavene.



Oppgave 1 Beskriv bildet (skriv mellom 80 og 100 ord)

Hva ser du på bildet? Hva gjør personene?

Oppgave 2 Skriv en e-post til en venn. Skriv mellom 80 og 200 ord.

Fortell hva du gjorde sist sommer.

Oppgave 3 Skriv en e-post til skolen. Skriv minimum 80 ord.

Skolen skal flyttes til en by lengre fra deg. Dette skaper problemer for deg og flere av vennene dine. Skriv en e-post til skolen og forklar hvorfor det er viktig at skolen ikke skal flyttes.

Vedlegg 2: Prinsipper for feiltelling og korrigering av tekstene

A) **To hovedprinsipper** (fra Tenfjord, Hagen & Johansen, 2009, s. 60):

Det pragmatiske probabilitetsprinsippet

Velg den tolkningen av innholdet som er *mest sannsynlig* fra et pragmatisk synspunkt når man tar i betraktning teksten som helhet, i tillegg til den situasjonelle og kulturelle konteksten.

Det minimale modifikasjonsprinsippet

Velg den rekonstruksjon av teksten som utgjør *minst mulig endring* av originalen. Med andre ord: Velg det alternativet som medfører minst korreksjon.

B) **T-enheter, klaususer, feiltelling og ordtelling** (tilpasset norsk fra Polio 1997, Appendix C og Polio og Shea 2014, Appendix A):

Her listes bare det som forekommer i tekstene i undersøkelsen.

T-enheter

- En t-enhet defineres som en uavhengig klausus og alle avhengige klaususer.
- Setningsfragment (som mangler subjekt/finitt verbal) telles ikke som t-enhet, men henges på foregående t-enhet.
- Underordnet klausus som står alene henges på foregående klausus, og telles som en t-enhet med en feil.
- Hvis grammatisk subjekt som ikke kan sløyfes mangler i en sideordning, telles hele setningen som en t-enhet. Eks. * «Først vi gikk til Oslo, så reiste til Sverige.»[20]
- Etterhengte taggspørsmål telles ikke som en t-enhet, men henges på foregående t-enhet.

Klaususer

- En klausus inneholder et subjekt og et finitt verb.

[20] En asterisk (*) brukes i språkvitenskapelige sammenhenger for å markere ugrammatiske eller målspråksavvikende setninger eller former (Abrahamsson, 2009, s. 22).

- Imperativ trenger ikke subjekt for å telles som klausus.
- En setning med subjekt og bare hjelpeverb regnes ikke som egen klausus eller t-enhet. Eks. * «Jeg hadde ferie og vennen min ville også».

Feiltelling

- Én feilaktig, manglende eller ekstra bokstav telles ikke som feil.
- Omkast av bokstaver i et ord telles som én feil. Eks. * «Vi tok fsik i vannet…»
- Kommafeil telles ikke som feil.
- Ekstra eller manglende stor bokstav telles som feil.
- Manglende punktum telles ikke som feil hvis stor bokstav kommer umiddelbart etter.
- Ekstra punktum telles bare som feil hvis det også er stor bokstav umiddelbart etter.
- Grammatisk referanse telles som én feil hvis det er tvil om hva som er korrelatet.
- Feil som kan gjøres av morsmålsbrukere telles også som feil (for eksempel da/når, og/å, feil bruk av apostrof ved genitivs-s, orddeling).
- Feilaktig samsvarsbøying telles bare som én feil. Eks. * «…min ny møbler.»
- Feil register eller stilnivå telles ikke som feil, heller ikke inkonsekvent bruk.

Ordtelling

- Tall telles som ett ord.
- Egennavn telles som de er skrevet.
- Ord med bindestrek telles som ett ord

Vedlegg 3: Forespørsel til rektor om deltakelse i prosjektet

Forespørsel om deltakelse i forskningsprosjektet

«Voksnes skriving på et andrespråk»

Bakgrunn og formål

Jeg er student ved Høgskolen i Hedmark og skal i et masterprosjekt gjennomføre en undersøkelse om voksnes skriving på et andrespråk. Jeg vil sammenligne skriving med penn på papir og skriving på datamaskin.

Jeg ønsker å bruke tre klasser ved skolen fordi de har et godt nok norsknivå. Elevene i klassene får hvert sitt skjema som tilsvarer dette og signerer hvis de ønsker å delta.

Hva innebærer deltakelse i studien?

Klassene deles i to og halvparten av hver klasse får en skriveprøve på papir, mens den andre halvparten får samme skriveprøve på datamaskin. Elevene vil også få et spørreskjema der jeg vil vite litt om alder, kjønn, hvilket land de kommer fra, hvor mange år de har bodd i Norge, hvilken utdanning de har og omtrent hvor mange norsktimer de har hatt.

Hva skjer med informasjonen?

Alle personopplysninger vil bli behandlet konfidensielt. Bare jeg og mine veiledere vil ha tilgang til personopplysninger. Nøkkelen som kobler elevene med deres opplysninger finnes bare i papir og ligger innelåst hjemme hos meg. Det vil ikke bli mulig å kjenne igjen enkeltpersoner i den ferdige oppgaven.

Prosjektet skal etter planen avsluttes 15. mai 2016. Personopplysningene blir da anonymisert og koblingsnøkkelen blir makulert.

Frivillig deltakelse

Det er frivillig å delta i studien, og eleven kan når som helst trekke sitt samtykke uten å oppgi noen grunn. Dersom vedkommende trekker seg, vil alle opplysninger om eleven bli anonymisert.

Dersom du har spørsmål til studien, ta kontakt med Pål-Otto Mikkelsen, tlf. 47 31 53 61, epost: paal-mi@online.no eller min hovedveileder Marte Monsen, tlf. 62 51 72 33, epost: marte.monsen@hihm.no.

Studien er meldt til Personvernombudet for forskning, Norsk samfunnsvitenskapelig datatjeneste AS.

Samtykke til deltakelse i studien

Jeg har mottatt informasjon om studien, og er villig til å la skolen delta

---

(Signert av rektor, dato)

Vedlegg 4: Forespørsel til elever om deltakelse i prosjektet

Forespørsel om deltakelse i forskningsprosjektet

«Voksnes skriving på et andrespråk»

Bakgrunn og formål

Jeg heter Pål-Otto Mikkelsen og er til daglig lærer ved (sted) Læringssenter. Jeg er også student ved Høgskolen i Hedmark og skal i et masterprosjekt gjennomføre en undersøkelse om voksnes skriving på et andrespråk. Jeg vil sammenligne skriving med penn på papir og skriving på datamaskin.

Jeg ønsker å ha deg med i denne studien fordi du har lært norsk en stund. Alle i din klasse og en annen klasse ved skolen blir spurt om å delta.

Hva innebærer deltakelse i studien?

Klassene deles i to og halvparten av hver klasse får en skriveprøve på papir, mens den andre halvparten får samme skriveprøve på datamaskin. Du vil også få et spørreskjema der jeg vil vite litt om alder, kjønn, hvilket land du kommer fra, hvor mange år du har bodd i Norge, hvilken utdanning du har og omtrent hvor mange norsktimer du har hatt.

Hva skjer med informasjonen om deg?

Alle personopplysninger vil bli behandlet konfidensielt. Bare jeg og mine veiledere vil ha tilgang til personopplysninger. Nøkkelen som kobler deg med dine opplysninger finnes bare i papir og ligger innelåst hjemme hos meg. Det vil ikke bli mulig å kjenne igjen enkeltpersoner i den ferdige oppgaven.

Prosjektet skal etter planen avsluttes 15. mai 2016. Personopplysningene blir da anonymisert og koblingsnøkkelen blir makulert.

Frivillig deltakelse

Det er frivillig å delta i studien, og du kan når som helst trekke ditt samtykke uten å oppgi noen grunn. Dersom du trekker deg, vil alle opplysninger om deg bli anonymisert.

Dersom du har spørsmål til studien, ta kontakt med Pål-Otto Mikkelsen, tlf. 47 31 53 61, epost: paal-mi@online.no eller min hovedveileder Marte Monsen, tlf. 62 51 72 33, epost: marte.monsen@hihm.no.

Studien er meldt til Personvernombudet for forskning, Norsk samfunnsvitenskapelig datatjeneste AS.

Samtykke til deltakelse i studien

Jeg har mottatt informasjon om studien, og er villig til å delta

---

(Signert av prosjektdeltaker, dato)

Vedlegg 5: Inndeling av ordklasser i funksjonsord og innholdsord

Norsk referansegrammatikk (NRG) (Faarlund, Lie & Vannebo, 1997, s. 21) skiller mellom *grammatiske ord*, som angir relasjoner innenfor språket selv, og *leksikalske ord*, som refererer til noe «i verden», som *hest* og *løpe*. (NRG har også med *pro-ord*, som henter sin betydning fra konteksten eller situasjonen. Dette skillet er ikke hensiktsmessig i min sammenheng, og jeg behandler dem sammen med de øvrige ordene. I praksis betyr det at de må vurderes i hvert enkelt tilfelle). Grammatiske ord tilsvarer funksjonsord, og leksikalske ord tilsvarer innholdsord. Skillet mellom ord som har en *referent*, og ord som ikke har det, er mitt hovedskille mellom funksjonsord og innholdsord. Semantikken skiller mellom ulike typer referanser (Saeed, 2003), men det er ikke relevant her. Noen ordklasser er enten funksjonsord eller innholdsord, mens andre må vurderes i hvert enkelt tilfelle, dette er vist i tabell 15. I midtre spalte vises klassifiseringen i NRG, i høyre spalte mine vurderinger delvis basert på framstillingen i Golden (2014):

Tabell 15. Ordklasser, semantisk type og vurdering

Ordklasse	**Semantisk type (etter NRG)**	**Vurderes enkeltord?** (delvis etter Golden, 2014, s. 41ff)
Substantiv	Leksikalsk	Nei, ordene har referent eller får sin betydning i konteksten.
Verb	Leksikalsk	Nei, ordene har referent eller får sin betydning i konteksten. Hjelpeverb er grammatiske ord.
Adjektiv	Leksikalsk	Ja
Pronomen	Pro-ord	Ja
Determinativ	Pro-ord	Ja
Preposisjon	Leksikalsk eller grammatisk	Ja
Adverb	Leksikalsk	Ja
Subjunksjon	Grammatisk	Nei
Konjunksjon	Grammatisk	Nei
Interjeksjon	Ingen	Ja, noen ord kan ha en referent (*faen, pokker* osv.*)*

Vi ser at ordene vurderes enkeltvis i seks av ordklassene. Det var færre enn ti interjeksjoner totalt i alle mine tekster.