Leksikosyntaktiske trekk og skriveverktøy

En kvantitativ undersøkelse av tekster skrevet for hånd og på tastatur av elever i VG1

Bård Uri Jensen

Avhandling for graden philosophiae doctor (ph.d.)

ved Universitetet i Bergen

2017

Dato for disputas: 5. desember 2017

År: 2017

Tittel: Leksikosyntaktiske trekk og skriveverktøy

En kvantitativ undersøkelse av tekster skrevet for hånd og på tastatur av elever i VG1

Forfatter: Bård Uri Jensen

Trykk: AiT Bjerch AS / Universitetet i Bergen

Fagmiljø

Arbeidet med denne avhandlingen har vært tilknyttet Forskerskolen i språkvitenskap og filologi ved Det humanistiske fakultet og Institutt for lingvistiske, litterære og estetiske studier ved Universitetet i Bergen. Kandidaten har vært ansatt ved Høgskolen i Hedmark, Avdeling for lærerutdanning og naturvitenskap.

Forord

Denne avhandlingen representerer et arbeid som har tatt mange år. Mange fortjener takk for sitt bidrag i denne prosessen.

Først og fremst må jeg varmt takke hver og en av mine tre veiledere: min hovedveileder gjennom hele perioden, Torodd Kinn, min biveileder gjennom hele perioden, Lars Anders Kulbrandstad, og Gard Buen Jenset, som kom inn som ekstra biveileder litt senere i prosessen. Alle har lest umenneskelige mengder med tekst og bidratt uvurderlig til å forme både meg og prosjektet.

Ved det som nå heter Uni Research i Bergen er det flere som fortjener takk; først og fremst Paul Meurer, som gjorde tilpasninger til korpusløsningene slik at de kunne løse mine behov, og dessuten har svart på mange spørsmål undervegs, men også Knut Hofland, som har bidratt med innspill og i diskusjoner, og daværende direktør Gisle Andersen, som var den som først foreslo at jeg skulle bruke deres korpusløsning til min tekstsamling. Uten dette initiativet ville prosjektet ha sett svært annerledes ut.

Jeg vil også takke Kari Tenfjord for å introdusere meg til ASK, Norsk Andrespråkskorpus, og for senere å invitere meg og inkludere meg i mange spennende aktiviteter knyttet til dette prosjektet og det senere Askeladden-prosjektet. Denne kontakten har vært viktig for utviklingen av både prosjektet mitt spesielt og kompetansen min mer generelt, og det er mange personer tilknyttet ASK og Askeladden som fortjener en takk for nyttige kommentarer, spørsmål og diskusjoner undervegs.

Også ved Tekstlaboratoriet ved Universitetet i Oslo er det mange som fortjener takk. Janne Bondi Johannessen leste en tidlig versjon av prosjektbeskrivelsen min og kom med mange og nyttige tilbakemeldinger. Anders Nøklestad, Kristin Hagen og Lars Nygaard har alle ytt ulike former for viktig teknisk støtte, og mange andre medarbeidere har svart på spørsmål.

Alle bidragsytere kan ikke nevnes ved navn. Jeg må takke fire anonyme lærere i videregående skole. De har vist en sportslig innstilling og utført ikke ubetydelig merarbeid for å samle inn tekster og andre data. En spesiell takk til den fagansvarlige læreren som tok initiativet til at nettopp deres skole kunne hjelpe meg med datainnsamlingen. Også alle de anonyme elevene som bidro med tekster og fylte ut spørreskjema skal ha takk; uten dem inget prosjekt.

Mange konferanse- og seminardeltagere har bidratt til prosjektet gjennom tilbakemeldinger til presentasjoner på konferanser og seminarer. Kolleger, venner og mer tilfeldige bekjentskaper har bidratt gjennom uformelle samtaler, og studenter har bidratt gjennom spørsmål og diskusjoner i undervisningen. Alle fortjener en takk, men alle kan ikke nevnes her.

Mange har bidratt til at prosjektet fremstår slik det gjør i denne avhandlingen, men alle feil og mangler er selvfølgelig mitt ansvar.

Til slutt må jeg takke min arbeidsgiver Høgskolen i Hedmark (fra 1. januar 2017 Høgskolen i Innlandet), som tildelte meg doktorgradsstipend og dermed gjorde det økonomisk mulig å gjennomføre prosjektet.

Abstract

This PhD thesis investigates lexicosyntactic features in writings in Norwegian L1 by 60 16-year-old pupils written using two different writing tools: hand-writing and typing. Its methodological approach is corpus-based and statistical, and its theoretical foundations are mainly those of theories of complexity and register variation.

The main focus of the thesis is on the analysis of 5 lexical and 8 syntactic variables. The lexical variables are average word-length, average word-length in lexical words, lexical density, global TTR and local TTR, both adjusted or neutralised for text-length. The syntactic variables are t-unit length, clause length, frequency of short subclauses, number of prepositional phrases per clause, number of adverbial subclauses per clause, number of subclauses per t-unit, ratio of t-units with a short frontal constituent and frequency of attributive adjectives. All these are analysed using anova on four pupil features, which are gender, general writing skills, total text length of the two texts and ratio of text lengths in the two texts. Ten of the variables are included in an overall principal component analysis.

The main findings come in four general categories:

Some variables seem unaffected by the writing tool. This applies to average clause length and frequency of attributive adjectives.

Some variables display a shift in the direction of more spontaneous, "oral" features in the typed texts. This applies to subclause frequency and perhaps average t-unit length.

Some variables have different properties in boys' and girls' writings; they display more spontaneous features in the typed texts written by girls and more planned or edited features in the typed texts written by boys. This applies to average word-length, average word-length in lexical words, global TTR and ratio of short frontal constituents.

Some variables have different properties in the writings of pupils who write considerably longer texts when typing compared to the writings of pupils who write texts of approximately equal lengths with both writing tools. The former pupils have more planned or edited features in their typed texts, whereas the latter pupils have more spontaneous features in their typed texts. This applies to local TTR, global TTR and ratio of adverbial clauses.

The principal component analysis confirms the two effects splitting the sample of pupils in terms of gender and production length.

In addition, some variables display types of patterns which do not fit neatly into any of the four categories above, namely average t-unit length (longer in typed texts), ratio of short subclauses (complex interactions of several pupil-related factors), lexical density (interactions of two pupil-related factors), frequency of prepositional phrases (higher in typed texts by productive writers, lower in typed texts by terse writers), and ratio of short frontal constituents (interactions between difference in text length and gender).

An important goal of the project was the further development of methods for this kind of study based on statistical stylistics. As part of the study, several lexical variables are constructed, compared and evaluated with respect to reliability and validity, among these entropy-based lexical distribution measures and frequency-based measures termed logarithmic frequency index. Also, emphasis has been put on investigating what relevant and valid textual or linguistic features are possible to extract from a corpus using a combination of automatic procedures and a limited extent of manual procedures, namely the manual segmentation of the texts into t-units and clauses, and correction of orthography and punctuation.

Innhold

1 Innledning

1.1 Bakgrunn

Tema for denne avhandlingen er leksikosyntaktiske trekk i elevers skriving med to ulike verktøy: på pc-tastatur med tekstbehandlingsverktøy og for hånd med penn eller blyant på papir. Det konkrete forskningsobjektet er 120 tekster som er produsert av 60 elever i første trinn på videregående skole (VG1), hver tekst i en dobbelt norsktime på skolen. Prosjektet har undersøkt monologiske tekster som er skrevet i en sakprosasjanger, som er lengre enn en viss minimumslengde. Tekstene er skrevet av elever som har norsk som førstespråk, og som er antatt å beherske tekstbehandlingsverktøyet.

Hovedproblemstillingen er hvorvidt og eventuelt hvordan leksikosyntaktiske trekk i elevtekster påvirkes av skriveverktøy, altså av om de er skrevet på tastatur eller for hånd.

Hypotesen jeg har hatt som utgangspunkt, er at endringer i ulike faktorer påvirker skriveprosessen og dermed produktet. Prosjektet støtter seg på empiri og teori om språklige variabler knyttet til kompleksitet, register og skriving, og de viktigste delhypotesene er at faktoren *økt skrivehastighet* vil påvirke tekstproduktene i retning av et spontant, prototypisk muntlig stilnivå, mens faktoren *bedre redigeringsmuligheter* vil påvirke tekstproduktene i retning av et planlagt, prototypisk skriftlig stilnivå. Andre variabler i skrivesituasjonen, som for eksempel oppgaveteksten og elevens skriveferdigheter og motivasjon, vil bestemme hvilken faktor som får størst gjennomslagskraft i hver tekst. Hypotesen blir presentert og diskutert mer inngående i kapittel 5, men kan sammenfattes slik:

> I skrivesituasjoner der hastighetsfaktoren har størst gjennomslag, vil tastetekster ha flere spontane ("muntlige") trekk enn de håndskrevne, mens i skrivesituasjoner der redigeringsfaktoren har størst gjennomslag, vil tastetekster ha flere planlagte ("skriftlige") trekk enn de håndskrevne.

Undersøkelsen benytter kvantitative metoder for å sammenligne fordelingen av leksikosyntaktiske trekk i et innsamlet korpus av elevtekster og sammenholde resultatene med faktorer i skrivesituasjonen, inkludert variabler knyttet til egenskaper ved elevene.

At datamaskinstøttet kommunikasjon er i ferd med å endre menneskers forhold til språk og språkbruk, er det liten tvil om, slik teknologiske nyvinninger som skrift, trykkekunst og telefon har gjort før. Ulike faktorer ved ytringssituasjonen kan potensielt påvirke språkbruken på ulike måter, og med dette prosjektet ønsker jeg å isolere faktorer som har med de psykolingvistiske prosesseringsbetingelsene i tekstbehandling å gjøre, og se på hvordan disse påvirker språkbruk.

1.2 Formål og begrunnelse

Formålet med prosjektet er knyttet til både resultater og metodeutvikling. Nytteverdien kan oppsummeres i følgende fem punkter:

1. Resultatene belyser hvilken innvirkning visse fysiske rammefaktorer i en skrivesituasjon har på språklige variabler i ytringene. Dette vil øke vår kunnskap om språkproduksjon generelt.
2. Resultatene bidrar til å belyse hvordan vår digitale hverdag påvirker språkproduksjon. Hvis det er slik at unge mennesker som behersker begge skriveverktøyene i dag, produserer språk med ulike kjennetegn med de to verktøyene, indikerer dette at språkbruken i samfunnet generelt vil endre seg med digitaliseringen av samfunnet. Dette prosjektets datainnsamling var kanskje siste mulighet til å foreta synkrone, kontrollerte forsøk med informanter som har tilstrekkelige skriveferdigheter både for hånd og med tastatur.
3. Prosjektet er skrivedidaktisk viktig med hensyn til både skriveopplæring og vurdering. Russell og Haney (1997) mente for eksempel allerede for 20 år siden at pc-kyndige elevers skriveferdigheter blir undervurdert når elevene blir vurdert ut fra sine håndskrevne tekster. Generelt vil økt kunnskap om den innvirkning som skriveverktøyet har på tekstproduktet, være nyttig i skriveopplæringen.
4. Prosjektet har dessuten frembrakt mer detaljert kunnskap om leksikosyntaktiske trekk i elevtekster generelt, uavhengig av skriveverktøy, selv om dette i seg selv ikke var et konkret mål med prosjektet.
5. I tillegg til de resultatmessige begrunnelsene bidrar arbeidet og avhandlingen også med metodologisk utvikling, både når det gjelder konstruksjon og utnyttelse av denne typen halvautomatisk oppbygde tekstkorpus av innlærerskrivere, og når det gjelder matematiske og statistiske metoder for analyse av dataene.

1.3 Notasjon og konvensjoner i avhandlingen

Dette avsnittet gir en oversikt over notasjonskonvensjoner i avhandlingen.

1.3.1 Teksteksempler

Avhandlingen inneholder mange teksteksempler, hovedsakelig fra elevtekstene som er samlet inn, men i noen grad også konstruerte eksempler eller eksempler hentet fra faglitteraturen. De fleste eksemplene er på setningsnivå, men mange er også på ord- eller frasenivå. Når man gjengir språklige eksempler i lingvistiske arbeider, er det vanlig å bruke fonemklammer (/.../) eller grafemklammer (<...>). I dette arbeidet er imidlertid hverken fonemer eller grafemer normalt gjenstand for oppmerksomhet, ettersom studieobjektet først og fremst er leksemer, morfosyntaktiske ord eller ordsekvenser. Jeg har derfor valgt å bruke en modusnøytral markering av de språklige eksemplene i løpetekst, og for å kunne reservere kursiv til utheving eller til markering av tekst på andre språk enn norsk, har jeg valgt å bruke omvendte skråstreker til å markere språklige eksempler: \eksempel\. I de tilfeller der eksemplene er gjengitt som nummererte eksempler i egne avsnitt, er ikke skråstrekene nødvendige.

I mange tilfeller er det nødvendig å gjengi den syntaktiske strukturen i eksemplene. Disse er noen ganger gjengitt med koder som representerer syntaktiske enheter. I korpuset er sekvensene kodet ved hjelp av XML-koder som normalt ikke er gjengitt i avhandlingsteksten. Eksemplet under viser hvordan annoteringen av den syntaktiske strukturen er gjort i korpuset (3), og hvordan dette kan være gjengitt i avhandlingen (1) og (2), avhengig av hvilken informasjon om segmentet som er relevant å gjengi.

(1) {Jeg sa {jeg var sulten.}}
(2) {$_{T}$ Jeg sa {$_{Snom}$ jeg var sulten.}}
(3) <t-unit>Jeg sa <clause type="nominal">jeg var sulten.</clause></t-unit>

Kilden for eksemplet er normalt en elevtekst, og filnavnet til elevteksten er gjengitt i hakeparentes etter eksemplet, for eksempel [A1-210]. Siden tekstene er relativt korte, har jeg ikke funnet det nødvendig å gjengi linje- eller sidenummer. Filnavnet i eksemplet over viser at teksten er svar på oppgave A1, og at den er skrevet av elev 210. Appendiks D1 gir en oversikt som viser om hver tekst er en håndtekst eller tastetekst, og elevens kjønn og skriveferdighet. Når jeg har konstruert eksemplet selv, står mine initialer etter eksemplet: [BUJ].

1.3.2 Diagrammer

"Ikke slipp individene av syne," sier Anne Golden (2010, s. 114). Dette er en viktig påminnelse i forbindelse med statistiske analyser. Ett viktig redskap for å holde øye med individene er diagrammer; mange av diagrammene i denne avhandlingen trekker nettopp frem individene i de tendensene som beskrives.

Hvert diagram i avhandlingen er ledsaget av forklaring, men enkelte notasjonskonvensjoner er gjennomgående og så vanlige at de ikke alltid blir forklart i tilknytning til hvert diagram.

1. Når diagrammet skiller mellom håndtekster og tastetekster, er håndtekstene markert med svart og tastetekstene med rødt. Markeringer som gjelder begge verktøy er ofte markert med grått, men kan også være gjengitt med svart der det ikke er fare for misforståelse.
2. Gutter er tegnet med lyseblått og jenter med rosa. (Av tekniske grunner kan den rosa fargen i visse sammenhenger på trykk framstå som mer lilla enn rosa.)
3. Sterke elever er kraftigere markert enn middels elever, for eksempel ved hjelp av fylte punkter eller fete linjer i motsetning til åpne punkter og tynne linjer, eller med mørkere fyllfarge.
4. Middelverdier eller medianer er gjengitt med stiplede linjer, mens regresjonslinjer eller -kurver er gjengitt med heltrukne linjer eller kurver.

Diagrammene i avhandlingen er generert av statistikkprogrammet R (R Core Team, 2016). I enkelte spesialiserte diagrammer ville det være uforholdsmessig arbeidskrevende å oversette diagramtekstene fra engelsk til norsk, ettersom tekstene er generert av programkode som ikke er enkelt tilgjengelig. I slike tilfeller står diagrammene med engelsk tekst og engelsk desimalpunkt, men i de fleste diagrammene er det norsk tekst og norsk desimalkomma.

1.4 Avhandlingens oppbygning

Foruten dette innledningskapitlet består avhandlingen av tre hoveddeler i tillegg til et avslutningskapittel.

Den første delen presenterer prosjektets teoretiske fundament. Hovedvekten ligger i et kapittel om språklig kompleksitet. Deretter følger et kapittel om språklig variasjon knyttet til register, skriving og modus. Til slutt i denne delen står et kort kapittel som presenterer grunnleggende lingvistiske begreper og enheter, før et eget kapittel går dypere inn i en drøfting av prosjektets hovedhypotese og forskningsspørsmål.

Den andre delen presenterer prosjektets metodiske innretning. Det første kapitlet i denne delen handler om datainnsamlingen, mens det andre kapitlet beskriver og drøfter metodene for statistisk analyse. I dette kapitlet går jeg blant annet ganske detaljert inn i en diskusjon om logit-transformering.

Den tredje delen presenterer analyser av datamaterialet og resultater av analysene. Først kommer et kapittel som presenterer viktige egenskaper ved henholdsvis elevene og tekstene i materialet. Deretter følger tre kapitler med hvert sitt fokus på forskjeller i språklige variabler mellom håndskrevne og tastede tekster. Det første av disse tre kapitlene dreier seg om informasjonell tetthet knyttet til ulike leksikalske variabler. Det andre handler om leksikalsk variasjon og det tredje om syntaktiske variabler. De to kapitlene om leksikalske variabler danner hovedtyngden i analysene, mens kapitlet om syntaktiske variabler kompletterer bildet ved å representere en annen og delvis uavhengig kompleksitetsdimensjon, slik jeg argumenterer for i teoridelen. Til slutt i avhandlingens tredje del forsøker jeg å danne et mer helhetlig bilde av variasjonen og kompleksiteten i materialet gjennom en multivariat prinsipalkomponentanalyse.

Gjennom de fem analysekapitlene går jeg dessuten mer detaljert inn i den teoretiske bakgrunnen og motivasjonen for de enkelte variablene, og jeg utvikler og drøfter metodiske grep som er lettest å beskrive og drøfte i sammenheng med de konkrete analysene.

Helt til slutt i teksten følger et kapittel som oppsummerer og konkluderer om avhandlingens bidrag til forskningen.

I avhandlingens appendiks er ulik dokumentasjon av metode og prosedyre gjengitt. Dette dreier seg om dokumentasjon av statistiske analyser som det ikke er naturlig å inkludere i avhandlingens brødtekst (A), ulike dokumenter knyttet til innsamling av tekster og data (B), noe tekniske data knyttet til oppbyggingen av korpuset (C), de konkrete tallene som er hentet ut av korpuset og som ligger til grunn for de faktiske analysene (D), og noe av programkoden som er benyttet i de statistiske analysene (E). Jeg har ikke funnet det naturlig å inkludere all programkoden som er brukt, ettersom dette ville øke avhandlingens omfang betydelig. Jeg har heller ikke inkludert elevenes tekster; dette er av sensitivitetshensyn. Sist blant appendiksene står et emneregister (F) og litteraturliste (G).

Teori

De tre neste kapitlene danner det teoretiske grunnlaget for prosjektet. Det første kapitlet utforsker ulike begreper for språklig kompleksitet. Det neste kapitlet handler om skriving, register, modus og språklig variasjon. Det tredje kapitlet presenterer grunnleggende språkvitenskapelige begreper og enheter, særlig syntaktiske.

Til slutt i denne delen utdyper jeg hypotesen og forskningsspørsmålene.

2 Språklig kompleksitet

Språklig kompleksitet har relevans for dette prosjektet fordi den overordnede hypotesen, slik den er presentert i avsnittene 1.1 og 5.1, er knyttet til at de ulike fysiske rammene for de to produksjonssituasjonene påvirker ytringenes form, og at påvirkningen skjer på en slik måte at den raskere produksjonsmåten – altså tastaturet – tenderer til å produsere høyere frekvens av språkformer som vi normalt knytter til spontan språkproduksjon, altså prototypisk tale. Det virker nærliggende at dette i så fall på en eller annen måte er relatert til kompleksitet i ytringen.

I dette kapitlet forsøker jeg å sirkle inn kompleksitetsbegreper som vil kunne operasjonaliseres til et slikt formål.

Jeg begynner med å presentere noen begreper fra allmenn kompleksitetsteori, slik den arter seg i matematikk eller naturvitenskap. Deretter vil jeg se på hvordan slike allmenne begreper kan benyttes og blir benyttet i lingvistikken, først innenfor kompleksitet i språksystemet, eller systemisk kompleksitet, og deretter i det jeg kaller strukturell kompleksitet, altså kompleksitet knyttet til ytringer, både på t-enhetsnivå og på tekstnivå. Dette vil jeg til slutt i kapitlet knytte til komputasjonell kompleksitetsteori i lingvistisk perspektiv.

2.1 Noen begreper fra allmenn kompleksitetsteori

Først skal jeg se på noen anvendbare begreper fra informasjonsteori og allmenn kompleksitetsteori. Jeg begynner med å diskutere kompleksiteten av strenger og mengder.

2.1.1 Kolmogorov-kompleksitet

Dette avsnittet baserer seg på Arecchi (2001), Dahl (2004), Gammerman & Vovk (1999) og Shen (1999).

Den aller enkleste innfallsvinkelen til kompleksiteten av en streng eller en mengde er relatert til strengens lengde eller en opplisting av elementene i mengden, altså eksemplarinventaret. En lengre streng eller en større mengde resulterer i en lengre opplisting, som den mest naive beskrivelsen av strengen eller mengden.

Imidlertid vil en beskrivelse av en streng eller en mengde kunne være kortere enn strengen eller mengden dersom det finnes mønstre eller regelmessigheter i dataene. Regelmessighetene kan utnyttes til å komprimere beskrivelsen. For eksempel vil strengen i (4) kunne beskrives som (aabb){8}, mens strengen i (5), som er en helt tilfeldig sekvens av *a* og *b*, vanskeligere kan beskrives på noen kortere måte enn ved å gjengi hele strengen.

(4) aabbaabbaabbaabbaabbaabbaabbaabb
(5) babbaabbabaabaabbbbbaabbabbabaaa

En beskrivelse av (4) kan altså gjøres kortere enn en beskrivelse av (5), og (5) kan dermed sies å være mer kompleks enn (4). Det er riktignok lokalt komprimeringspotensial også i (5);

for eksempel gjentas sekvensen *abb* fem ganger og kan representeres av et kortere symbol. Etableringsomkostningene ved et komprimeringsregime gjør det imidlertid vanskelig å oppnå reell komprimering i en så kort streng. I en lengre streng vil også tilfeldige sekvenser kunne komprimeres ved å representere frekvente substrenger med kortere symboler. Egentlig forutsetter teorien at den kortere beskrivelsen skal gjøres med det samme alfabetet som den opprinnelige strengen. Forklaringen i dette avsnittet tar dermed en snarvei, men den forklarer uansett de informasjonsteoretiske *prinsippene*.

Lengden av den kortest mulige beskrivelsen av en streng eller en mengde kalles strengens eller mengdens *Kolmogorov-kompleksitet* eller *informasjonell kompleksitet*.

2.1.2 Entropi og orden

Entropi er opprinnelig et termodynamisk begrep. I klassisk termodynamikk er is som smelter i vann, et eksempel på stigende entropi, men parallellen til informasjonsteori er klarere når man tar utgangspunkt i statistisk termodynamikk. I statistisk termodynamikk beskriver entropi grad av orden. Høy grad av orden – for eksempel olje som flyter i vann – beskrives av lav entropi, mens høy entropi – for eksempel en blanding av vann og alkohol – representerer uorden. Begrepet er senere overtatt av informasjonsteorien (Shannon, 1948, s. 10-), som også er et mer naturlig utgangspunkt for en diskusjon om språklig kompleksitet. I informasjonsteorien representerer høy entropi stor spredning eller varians i en mengde av meldinger, altså liten forutsigbarhet og høyt informasjonsinnhold. Lav entropi betyr stor forutsigbarhet, altså lite informasjonsinnhold. Entropien H(x) i en mengde av meldinger x regnes ut ved å summere produktene av hver meldings sannsynlighet og logaritmen av denne sannsynligheten:

$$(6) \quad H(x) = -\sum p(x_i) * \log(p(x_i))$$

Her står $p(x_i)$ for sannsynligheten for at meldingen x_i skal inntreffe. I informasjonsteoretiske sammenhenger brukes ofte logaritmebase 2, men dette er ikke nødvendig. Siden p-verdiene ligger mellom 0 og 1, og logaritmen av slike verdier er negativ, multipliseres summen med –1 for å oppnå en positiv verdi for entropien.

Et mål for effektivitet – en slags standardisert entropi E(x) – får vi ved å dividere entropien med logaritmen av antall meldinger:

$$(7) \quad E(x) = H(x) / \log(n)$$

E vil ha verdier mellom 0 og 1, der 0 representerer maksimal orden og 1 representerer maksimal uorden, og vil være uavhengig av antall mulige meldinger, i motsetning til entropien, som normalt vil øke med antall mulige meldinger. Samme logaritmebase må brukes i de to formlene.

Lovász (1997, s. 5-6) viser at man aldri kan vite om man har funnet den kortest mulige beskrivelse av en mengde, altså mengdens informasjonelle kompleksitet (2.1.1). Imidlertid kan man ved hjelp av entropien regne ut en grenseverdi nedad for en mengdes

informasjonelle kompleksitet, altså et teoretisk minimumsmål for den korteste beskrivelsen. Forholdstallet mellom det teoretiske minimumsmålet og den opprinnelige strengens faktiske lengde kaller Shannon for *relativ entropi*.

> The ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols will be called its relative entropy. This is the maximum compression possible when we encode into the same alphabet. One minus the relative entropy is the redundancy. The redundancy of ordinary English, not considering statistical structure over greater distances than about eight letters, is roughly 50%. (Shannon, 1948, s. 14)

Formelen for entropi viser dessuten at dersom vi skal beskrive en streng eller en mengde så kort som mulig, er det viktig å representere frekvente enheter med korte beskrivelser, enten disse enhetene er enkeltsymboler eller de er sekvenser eller grupper av symboler. Det er denne teknikken – LZ-algoritmen (Ziv & Lembel, 1977) – som ligger til grunn for mange komprimeringsprogrammer som brukes på filer i moderne datasystemer (Wikipedia, 2017). I praksis oppnår ikke algoritmen komprimering ned mot grenseverdien fordi behandlingen alltid vil medføre omkostninger i form av indeksering og lignende. (Se også 9.1 om Zipfs funn av sammenheng mellom ordfrekvens og ordlengde.)

Gries (2009, s. 112) bruker entropi til å vurdere kompleksiteten i substantivfraser i en tekst. I teksten er det 300 substantivfraser; 100 bestemte artikler, 100 ubestemte artikler og 100 fraser uten determinativ representerer total uorden, mens 300 determinativløse fraser representerer total orden. I det første eksemplet blir $E(x) = 1$, mens i det siste går $E(x)$ mot 0. (Entropien er udefinert dersom $p(x_i)=0$ for en av x_i.) Det første eksemplet er med andre ord mest komplekst.

2.1.3 Komputasjonell kompleksitet

La oss se på to nye eksempler på strenger dannet over samme alfabet:

(8) 12345678901234567890123456789012345678901234567890
(9) 14142135623730950488016887242096980785696718753769

(8) har et tydelig repeterende mønster av de ti sifrene i rekkefølge, mens (9) ser ut til å være en tilfeldig følge av siffer. (9) er imidlertid ikke tilfeldig; det er de 50 første sifrene i kvadratroten av 2. De to strengene har dermed informasjonell kompleksitet av samme størrelsesorden; faktisk har kanskje (9) den korteste beskrivelsen og dermed den laveste informasjonelle kompleksiteten. Dette virker kontraintuitivt, og det finnes et annet kompleksitetsbegrep som gjenspeiler inntrykket av at (9) er mest kompleks, nemlig *komputasjonell kompleksitet*. Komputasjonell kompleksitet (Lovász, 1997, s. 6) er relatert til spørsmål av typen "Hvor lang tid vil det ta å regne ut det *n*-te leddet i en streng?" Denne formen for kompleksitet kan ikke måles med en enkeltverdi, men vil være en funksjon av *n*. For (8) er det svært enkelt å regne ut det *n*-te leddet, siden svaret blir det siste sifferet i *n*. Man vet altså svaret i det øyeblikket man hører spørsmålet. Å regne ut det *n*-te leddet i (9), derimot, er mye vanskeligere og mer tidkrevende, og slik reflekterer den komputasjonelle kompleksiteten en intuitiv oppfatning av kompleksitet her.

2.1.4 Relativ kompleksitet

Östen Dahl (2004:24-25) peker med det han kaller for *relativ kompleksitet*, på at man i kompleksitetsstudier også må ta i betraktning premissene for en beskrivelse. Hvis man for eksempel skal beskrive en person, trenger man ikke bruke mye plass på å beskrive antall armer, ben, lunger, nyrer, etc., dersom personen er i tråd med et prototypisk menneske, eller "forholder seg til standarden". Bare ved avvik vil vi påpeke at en person for eksempel har bare én arm. Miestamo bruker betegnelsen relativ kompleksitet med en annen betydning enn Dahl, slik jeg forklarer i 2.2.1.

En strengs eller mengdes informasjonelle kompleksitet vil med andre ord bli påvirket av hvilken informasjon som blir tatt for gitt, eller hva slags teori en beskrivelse bygger på. En streng som avviker fra standarden i teorigrunnlaget, vil fremstå som mer kompleks enn en som forholder seg til standarden. Dette skal jeg komme tilbake til når jeg diskuterer entropi-basert kompleksitet i ytringer i 2.3.4. (Se også 10.5.2 og 11.1.3.)

2.1.5 Symmetri og asymmetri

I mange tilfeller er diskrepansen mellom informasjonell kompleksitet og komputasjonell kompleksitet relatert til at informasjonell kompleksitet ikke skiller mellom den korteste *beskrivelsen* av en streng og den korteste *definisjonen*. Definisjonen vår av (9) over genererer sekvensen, men før den er generert, vet vi lite eller ingenting om hvordan den ser ut. Beskrivelsen av (8), derimot, gir oss umiddelbart et bilde av hvordan strengen ser ut.

Mange problemer er ganske symmetriske når det gjelder komputasjonell kompleksitet. For eksempel er det omtrent like vanskelig å regne ut summen av to *n*-sifrede tall som å finne differansen mellom summen og ett av tallene. Imidlertid er mange problemer asymmetriske; for eksempel er det mye vanskeligere å finne kvadratroten av et kvadrattall enn å kvadrere et tall. Det finnes mange slike asymmetriske fenomener, for eksempel tredjegradsligninger, som er vanskelige å løse men lette å konstruere og regne ut verdien av. Det er for eksempel mye vanskeligere å løse ligningen i (10), som har løsningen x = 2, enn å regne ut verdien av y i (11) når x = 2.

(10) $x^3 + x^2 + x + 2 = 16$
(11) $y = x^3 + x^2 + x + 2 \quad | x = 2 (y=16)$

Slike asymmetriske fenomener benyttes i krypteringsalgoritmer. Det er lett å regne ut om passordet du skriver inn i et innloggingsvindu, stemmer med det krypterte passordet som er lagret i databasen, men det er svært tidkrevende å regne ut det ukrypterte passordet ut fra det krypterte.

2.2 Systemisk kompleksitet i lingvistikken

Jeg skal nå se på hvordan disse kompleksitetsbegrepene kan anvendes – og blir anvendt – i lingvistikken, og jeg begynner med det som Dahl (2009, s. 51) og Sampson (2009, s. 13) kaller *systemkompleksitet*, altså kompleksiteten i det grammatiske systemet.

Systemkompleksitet står dermed i motsetning til ytringskompleksitet – eller strukturell kompleksitet (Sampson, 2009, s. 13) – som jeg kommer tilbake til i 2.3, og som har mer direkte relevans for dette prosjektet.

Data fra kompleksitetsstudier kommer blant annet fra forskning på L1-utvikling og L2-innlæring. Teorien har relevans for disse emnene, selvfølgelig, men også for diakron språkvitenskap og for språkets fylogenese, altså utviklingen av menneskeartens språk og språkevne. Men mye av den oppblomstrende interessen for språklig systemkompleksitet de siste årene har dreid seg om språktypologi, særlig knyttet til den såkalte ALEC-hypotesen (*All Languages are of Equal Complexity*), som hevder at alle verdens språk har lik global systemkompleksitet. Nichols (2009, s. 121) peker på at ALEC-hypotesen neppe er den mest aksepterte hypotesen nå, i alle fall ikke blant yngre lingvister, og mange av artiklene i Sampson, Gil & Trudgill (2009) utfordrer og tilbakeviser ALEC-hypotesen.

2.2.1 Objektiv eller brukerorientert kompleksitet

Det er nødvendig å skille mellom objektiv, teoriorientert kompleksitet og subjektiv, brukerorientert kompleksitet (Dahl, 2009, s. 50-52; Miestamo, 2008, s. 24). I et teoriorientert perspektiv brukes utelukkende matematiske og informasjonsteoretiske modeller av den typen som er omtalt i 2.1.1 og 2.1.2 ovenfor, mens brukerorientert kompleksitet tar konkret prosesseringsmaskineri med i betraktning og er relatert til hvor krevende grammatikken eller språket er for et individ, i et lærings-, produksjons- eller resepsjonsperspektiv. Miestamo bruker betegnelsene absolutt og relativ kompleksitet i stedet for henholdsvis teoriorientert og brukerorientert, men jeg følger Dahls terminologi her og beholder betegnelsen *relativ kompleksitet* på et annet begrep (2.1.4).

Miestamo (2008, s. 26) peker på at det ikke er mulig å definere brukerorientert kompleksitet uten å ta hensyn til både brukergruppe og aktivitet. Han viser til at *kostnadene* i taleprosessen og i lytteprosessen, og for førstespråkstilegnere og andrespråksinnlærere, er ulike for forskjellige språklige fenomener. I mange sammenhenger, blant annet i dette prosjektet, må man dessuten ta hensyn til ulike kostnader knyttet til de ulike rollene *leser* og *skriver*. En generell teori om brukerrelatert kompleksitet vil måtte vekte de ulike rollenes vanskeligheter, og siden vi dessuten mangler mye kunnskap om hvor vanskelige ulike fenomener er i de ulike situasjoner (Miestamo, 2009, s. 82), trekker Miestamo (2008, s. 26-27) den konklusjon at vi foreløpig bør konsentrere oss om objektive kompleksitetsstudier.

Jeg skal derfor i første omgang være mest opptatt av teoriorientert kompleksitet, men vende tilbake til det brukerorienterte perspektivet når jeg drøfter kompleksitet i ytringer.

2.2.2 Effektiv kompleksitet

Östen Dahl (2009, s. 51) peker på det kontraintuitive i å anvende informasjonell kompleksitet på språksystemer og mener at *effektiv kompleksitet*, slik Gell-Mann (1995, s. 16-) definerer det, er et mer fruktbart kompleksitetsmål innenfor lingvistikken. Han begrunner dette med at maksimal *tilfeldighet* gir maksimal informasjonell kompleksitet,

mens det er kompleksiteten i lingvistiske *mønstre* vi er opptatt av, og det er denne typen kompleksitet i strukturer som fanges av begrepet effektiv kompleksitet. Miestamo (2009, s. 81) sier omtrent det samme: "[Kolmogorov complexity takes] total chaos as maximally complex. That is not the kind of complexity that interests us." Han sier videre at et effektivt kompleksitetsbegrep, som bare tar hensyn til regelmessigheter i systemet, er mer lingvistisk interessant.

Jeg har to litt ulike typer innvendinger mot dette. For det første virker det åpenbart intuitivt at et system med "huller", altså unntak eller uregelmessigheter, er mer komplekst enn et fullstendig regelmessig mønster. Dette er ikke i strid med informasjonell kompleksitet; en beskrivelse av et hullete mønster må bli lengre enn en beskrivelse av et fullstendig unntaksløst mønster. I forlengelsen av det må et mønster med mange huller være mer komplekst enn et mønster med få huller. Etter hvert som antall huller i et mønster stiger, nærmer vi oss fullstendig kaos eller tilfeldighet og dermed en beskrivelse som er like lang som systemet i seg selv. Ett eller annet sted på veien mellom unntaksløst mønster og fullstendig tilfeldighet vil en beskrivelse basert på mønster med unntak bli lengre enn en opplisting av inventaret.

Det er vanskelig å tenke seg et kompleksitetsmål for lingvistiske systemer som ikke tar hensyn til unntak fra regelmessigheter, og det er også vanskelig å tenke seg et kompleksitetsmål som setter en grense et sted på veien fra fullstendig systematikk til fullstendig kaos, slik at én modell skal brukes på den ene siden av grensen og en annen modell på den andre.

Dessuten er det jo heller ikke slik at informasjonell kompleksitet ikke tar kompleksiteten i systematikken med i beregningen; også ifølge informasjonell kompleksitet vil et mer omfattende mønster være mer komplekst enn et enklere mønster. Dette går også fram av et entropibasert mål; en mengde med flere enheter vil normalt få høyere entropi.

Den andre typen innvending er rettet mot det faktum at språk og grammatikk faktisk *er* overveiende systematiske. Det vil si at når vi diskuterer lingvistiske systemer, trenger vi i praksis ikke å ta hensyn til det tilsynelatende paradoks at en fullstendig kaotisk streng er mer kompleks enn en streng med visse regelmessigheter. Vi har i grammatikk *alltid* å gjøre med enheter som overveiende er bygget av systematikk. Dermed har det i praksis liten betydning om vi bruker informasjonell kompleksitet eller effektiv kompleksitet. Det er kanskje riktig å ta et lite forbehold her, på bakgrunn av grenseoppgangen mellom den kortest mulige beskrivelse og den kortest mulige definisjon, som jeg diskuterte tidligere (2.1.5). Dersom det skulle finnes en definisjon av en grammatikk som ville gjøre det nødvendig med omfattende kalkulasjoner for å generere grammatikkreglene, altså en definisjon av typen kvadratroten av 2, som vi så på tidligere, ville en slik svært komprimert beskrivelse, som altså ville representere den informasjonelle kompleksiteten, være lite relevant for grammatikken. Men det er ikke på dette grunnlaget Dahl og Miestamo kritiserer den informasjonelle kompleksitetens relevans for lingvistikk.

Dette er i tråd med Trudgills fremstilling om pidginisering ved kontakt (Trudgill, 2009, s. 100), selv om han ikke diskuterer informasjonell kontra effektiv kompleksitet eksplisitt. Han skriver at forenkling av språk består av tre hovedkomponenter:

- reduksjon i antall uregelmessigheter
- økning i leksikalsk og morfologisk gjennomsiktighet
- tap av redundans

Trudgill har som utgangspunkt (2009, s. 99) at kompleksitet kan knyttes til hvor vanskelig et andrespråk er å lære for ungdom og voksne, altså ett aspekt av brukerorientert kompleksitet. Trudgill peker på at forenklingsprosesser av typene over gjør slik læring og dermed språksystemene enklere. Dette synet er i hvert fall med hensyn til uregelmessigheter kompatibelt med informasjonell kompleksitet. Også Miestamo (2008, s. 29) og McWhorter (2008, s. 167) ser på uregelmessighet som et bidrag til systemkompleksitet. Dahl diskuterer ikke uregelmessigheter eksplisitt i denne sammenhengen.

Nichols (2009, s. 111-) peker på at ikke bare antall enheter i et system, men også den paradigmatiske variasjonsmuligheten for hver enhet, bidrar til kompleksitet. For eksempel vil et morfologisk system med mange allomorfer være mer komplekst enn et med få. Dette er i tråd med et entropi-basert kompleksitetsmål, og det virker også intuitivt riktig at en variabel med n+1 verdier er mer kompleks enn en variabel med n verdier, selv om en beskrivelse av n+1 ikke *trenger* å være lengre enn en beskrivelse av n. Et system med flere verdier kan være mindre komplekst dersom symmetrier i systemet bidrar til å økonomisere beskrivelsen.

2.2.3 Relativ kompleksitet

Jeg var i (2.1.4) innom Dahls begrep relativ kompleksitet. Relatert til relativ kompleksitet er hvordan de teoretiske rammene for en beskrivelse påvirker kompleksiteten i beskrivelsen ved å legge begrensninger på hvilken form beskrivelsen kan ta. I Chomsky-hierarkiet (Partee, ter Meulen, & Wall, 1993, s. 448-450) er det formalismen rundt frasestrukturreglene eller automatspesifikasjonen som er med og danner premisser for beskrivelsen. Chomsky-hierarkiet plasserer ulike formalismer på en skala over generativ kraft. Type-0-grammatikker, på toppen av skalaen, har størst generativ kraft og kan generere alle spesifiserbare språk, mens type-3-grammatikker, på bunnen av skalaen, har svakest generativ kraft og kan bare generere regulære språk. Type-2-grammatikker kan i tillegg til å generere regulære språk også generere kontekstfrie språk.

I (12) og (13) spesifiseres rammebetingelsene for henholdsvis regulære frasestrukturregler (type 3) og kontekstfrie frasestrukturregler (type 2) slik (Partee, et al., 1993), der den tomme streng er gjengitt med λ:

(12) $A \rightarrow x\psi \mid x \times \{a,b,c,...\}^{*} \psi \times \{A,B,C,..., \lambda\}$

(13) $A \rightarrow \varphi\chi\psi \mid \varphi\chi\psi \times \{a,b,c,...\ A,B,C,...\ \lambda\}$

Begge definisjonene er relativt enkle, og ingen av dem kan sies å være vesentlig mer kompleks enn den andre. Vi sier likevel gjerne at et språk som kan spesifiseres av en regulær grammatikk (type 3), er mindre komplekst enn et språk som krever en kontekstfri grammatikk (type 2), uavhengig av hvorvidt den resulterende grammatikken – i form av en mengde av frasestrukturregler – er like kompleks. Dette skyldes at språk med visse rekursive egenskaper ikke kan spesifiseres av den svakere formalismen.

Dette er altså ikke kompatibelt med informasjonell kompleksitet. For det første er ikke beskrivelsene av de premissgivende teoriene særlig ulike med hensyn til informasjonell kompleksitet. For det andre vil gjerne regulære frasestrukturregler for et (regulært) språk være mer informasjonelt komplekse enn kontekstfrie frasestrukturregler for det samme språket, på grunn av det kontekstfrie rammeverkets større uttrykkskraft, som ikke er eksplisitt uttrykt i (12) og (13). Generelt vil en grammatikk høyere opp i Chomsky-hierarkiet resultere i enklere grammatikkregler for samme språk, og dette er også blant Chomskys argumenter for å bruke transformasjoner – altså en type-0-grammatikk – selv om de kanskje ikke er formelt nødvendige (Chomsky, 1957, s. 49-60).

2.2.4 Regelkompleksitet og redundans

Jeg har så langt stort sett drøftet kompleksitet i lingvistiske inventar, altså for eksempel fonemsystemet eller et språks frasestrukturregler. Enkelhet i slike formale systemer kan ofte resultere i kompleksitet i regelsystemer som *mapper* et innholdssystem til et uttrykkssystem, for eksempel fra ord til skriftlig uttrykk eller fra argumentstruktur til syntaktisk form.

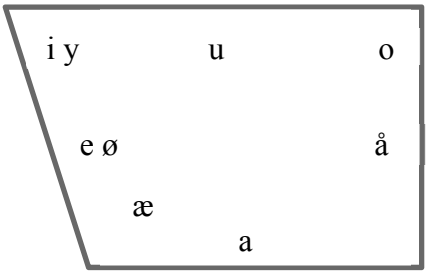
McWhorter (2001) forsøker å operasjonalisere et kompleksitetsbegrep ved å se spesielt på trekk som er obligatorisk markert, men ikke nødvendig for menneskelig kommunikasjon, og han trekker frem genus som et prototypisk eksempel (2001, s. 129). McWhorter og Trudgill kaller slike obligatoriske trekk redundans i språksystemet, selv om de i konkrete ytringer slett ikke nødvendigvis resulterer i redundans i informasjonsteoretisk forstand. Trudgill (2009, s. 100) skiller mellom to former for redundans av denne typen: morfologiske kategorier og (obligatorisk) dublering av informasjon i ytringer, for eksempel gjennom samsvarsbøyning. Genus i norsk er således et eksempel på begge typene, og et språksystem med genus og samsvarsbøyning av adjektiv, vil være mer komplekst enn et språk uten genus, som engelsk, om andre trekk er konstante. Den økte kompleksiteten skriver seg blant annet fra økt kompleksitet i regelsystemet som mapper fra innhold til uttrykk.

Dette er i tråd både med informasjonell kompleksitet og effektiv kompleksitet. En beskrivelse av et system med genus og samsvarsbøyning vil være lengre enn en uten, og ifølge Nichols (2009, s. 111-) vil også et genussystem med tre genus være mer komplekst enn ett med to når alt ellers er likt. Når det gjelder brukerrelatert kompleksitet, derimot, stiller det seg annerledes. McWhorter sier eksplisitt (McWhorter, 2001, s. 134) at han ikke hevder noen sammenheng mellom redundans og produksjons- eller prosesseringsvansker. Faktisk er det mye som taler for at redundans i forskjellige former *letter* resepsjonsprosessering (Nichols, 2009, s. 122).

2.2.5 Et enkelt eksempel

Et lingvistisk delsystem som illustrerer flere kompleksitetsbegreper godt, er alfabetet. Dette illustrerer også forskjellen mellom kompleksitet i beskrivelsen av inventaret og kompleksitet i beskrivelsen av regelsystemet.

Det norske vokalgrafeminventaret er mer komplekst enn det engelske; det består av 3 bokstaver mer, og selv om en symmetrisk mengde kan være mindre Kolmogorov-kompleks enn sin asymmetriske delmengde, er det vanskelig å se hvordan vokalgrafemene kan organiseres på en slik måte – enten man tar forholdet til vokalfoneminventaret med i betraktning eller ikke:



Figur 2-1: Grafeminventaret som brukes for vokalfonemer i norsk, tegnet inn i vokalfirkanten for å indikere deres primærfonem eller basisuttale (Jensen, 2005, s. 67, 74, 77)

Vi regner derfor med at nibokstavers-inventaret er mer komplekst enn seksbokstavers-inventaret[1], enten vi sammenligner inventaropplistinger eller mer komprimerte beskrivelser av inventaret, altså Kolmogorov-kompleksitet.

Når vi deretter går over til å se på regelsystemet i motsetning til inventaret, viser det seg imidlertid med en gang at det mest komplekse grafeminventaret gir vesentlig bedre økonomi enn det minst komplekse når det relateres til et (standard) østnorsk fonemsystem, ganske enkelt fordi det i stor grad gir en en-til-en-korrespondanse mellom grafemer og fonemer:

/i/ ↔ <i>, /y/ ↔ <y>, /o/ ↔ <å> (?), etc.

Et seksbokstaverssystem kunne gjøre bruk av digrafer med eksplisitte disambigueringsdiakritika, slik det gjøres i nederlandsk, eller man kunne akseptere flertydighet og rett og slett la noen grafemer ha dobbelt funksjon, slik det gjøres i italiensk. Begge løsninger ville øke kompleksiteten i regelsystemet i forhold til i et system med ett grafem for hvert fonem, enten gjennom et større antall regler, eller gjennom tvetydige regler, eller begge. Kompleksiteten i slike regelsystem er likevel bare marginalt høyere enn i et en-til-en-system, og dersom vi slår sammen inventarkompleksiteten og regelkompleksiteten i et globalt kompleksitetsmål, vil en vekting mellom de to delmålene være avgjørende for

[1] Fremstillinger av det engelske skriftsystemet klassifiserer ofte <y> som en konsonant, men bokstaven kan representere både vokalfonemer og konsonantfonemer, som i \lynx\ og \yet\.

resultatet. I et prosesseringsperspektiv er det likevel sannsynlig at seksbokstaverssystemet vil fremstå som mest komplekst, når det gjelder både produksjon og resepsjon.

La oss nå bygge videre på dette eksemplet og se på sammenhengen mellom informasjonell kompleksitet og entropi. Som kjent er bokstaven <o> et vanskelig område for norske skolebarn, ettersom bokstaven kan representere to fonemer, nemlig /u/ og /o/.[2] Dette er på alle måter et mer komplekst system enn om <o> alltid representerte /u/, og bare <å> ble brukt til å representere /o/. Så langt har vi ikke snakket om frekvens, men basert på fremstillingen over, kan vi tenke oss to prototypiske hypotetiske systemer:

a) Omtrent halvparten av <o>-ene representerer hver av /u/ og /o/.
b) Bare ett eller noen få av <o>-ene representerer det ene fonemet, for eksempel /o/; resten representerer /u/.

Vi forutsetter at det ikke er redundans i systemet av typen

(14) /o/ ~ <o> | $\_K_1K_1$

Slik redundans ville øke forutsigbarheten og redusere entropien. I det norske systemet er det stor grad av redundans av denne typen, så dette er en forutsetning som bryter med de faktiske forhold, men som forenkler eksemplet.

I tilfelle a) er entropien høy, med en effektivitet tilnærmet lik 1, mens i tilfelle b) er entropien lav, med en effektivitet som nærmer seg 0.

(15) $E_a = -\ (0{,}5*\log(0{,}5) + 0{,}5*\log(0{,}5)) / \log(2) = 1$
(16) $E_b = -(0{,}99*\log(0{,}99) + 0{,}01*\log(0{,}01)) / \log(2) = 0{,}081$

Graden av uorden – eller kompleksiteten – er altså høyest der grafemet representerer to nesten like store mengder. Dette betyr også at den informasjonelle kompleksiteten er høyest for a), mens potensialet for komprimering av regelsettet er høyest for b).

Situasjonen for norsk bokmål i dag vil jeg si ligner mest på det hypotetiske systemet a), selv om det selvfølgelig er en del regelmessigheter og symmetri som reduserer entropien. Til tross for slike mønstre fremstår området som komplekst og vanskelig å lære for både skolebarn og voksne, så i dette tilfellet er det godt samsvar mellom den teoretiske informasjonelle kompleksiteten og den intuitive oppfattelsen av kompleksiteten.

[2] Dette området kan og bør dessuten diskuteres i motsatt retning, altså at vokalen /o/ kan representeres av to forskjellige grafemer, men i dette illustrasjonseksemplet konsentrerer vi oss om retningen fra grafem til fonem, altså leseretningen. Eksemplet er også mer komplekst enn fremstillingen viser, ved at /u/ i tillegg kan representeres av <u>, men også dette ser vi bort fra her.

2.3 Kompleksitet i ytringer

Jeg skal nå se nærmere på kompleksitet i ytringer. Jeg vil begynne med å diskutere kompleksitet i setninger eller enkeltytringer, altså *strukturell kompleksitet*, som Sampson (2009) kaller det. Deretter vil jeg se på kompleksitet i lengre tekster.

2.3.1 Strukturell kompleksitet

De fleste arbeider om strukturell kompleksitet handler om "dybde" eller grad av klausal underordning. Dette er imidlertid ikke det eneste som bidrar til kompleksitet i ytringer. Michael Halliday sammenligner språklige mønstre i muntlige og skriftlige tekster og peker på at de har ulik "tekstur" (Halliday, 1987, s. 60). Bak metaforen skjuler det seg i hovedsak to relaterte leksikosyntaktiske variabler, nemlig grammatisk "innfløkthet" (*intricacy*) og leksikalsk tetthet. Når det gjelder grammatisk innfløkthet, peker Halliday på at varierte parataktiske og hypotaktiske forbindelser skaper kompleksiteten, men han ser altså også leksikalsk tetthet – en høy andel av leksikalske ord blant ordeksemplarene – som en form for kompleksitet. Jeg skal komme tilbake til leksikalsk kompleksitet når jeg omtaler kompleksitet på tekstnivå.

Jeg skal imidlertid først konsentrere meg om klausal dybde. De fleste er enige om at den slags kompleksitet i setningsstruktur må involvere hierarkisk struktur, altså trær. Sampson (2002) forsøkte å lage et mål på kompleksitet ved å måle hvert ords grad av klausal underordning. Miller og Chomsky (1963, s. 480-481) forsøkte å uttrykke kompleksitet som et forholdstall mellom ikke-terminale og terminale symboler i en setning. Dahl (2004, s. 104) relaterer kompleksiteten til antall derivasjoner for å generere den. Yngve (1961) forsøkte å knytte kompleksitet til antall venstreforgreninger i treet, mens mange har diskutert sentralinnføyde klaususer (Chipere, 2009; Karlsson, 2007; Miller & Isard, 1964). Disse arbeidene dreier seg i stor grad om regelkompleksitet, og de har alle et mer eller mindre komputasjonelt og brukerorientert fokus.

Alle knytter imidlertid kompleksiteten til trær, og dette er problematisk, ettersom det ikke finnes noen konsensusteori som bestemmer trestruktur. Dette var også Yngves poeng når han påpekte at hans egen dybdehypotese ikke er testbar (Yngve, 1998, s. 633-635; 2006, s. 9). Hvis vi legger dette ikke ubetydelige problem midlertidig til side, er det for så vidt ikke noe i veien for å analysere trærs kompleksitet i et ikke-komputasjonelt perspektiv. Et tres informasjonelle kompleksitet er lengden av den kortest mulige beskrivelsen av det – uavhengig av hvordan vi ser for oss at treet skal prosesseres. Det er imidlertid langt fra åpenbart hvordan en slik beskrivelse vil påvirkes av slike egenskaper ved trær som vi ellers er interessert i, for eksempel dybde versus bredde, høyrevridning versus venstrevridning, senterinnføying, etc. og det er mye som tyder på at trær først og fremst er interessante i et komputasjonelt perspektiv, selv om man i et sorteringsperspektiv vil kunne si at en ordning av et visst antall elementer i et dypt tre kunne representere større *orden* enn de samme elementene i et grunt tre.

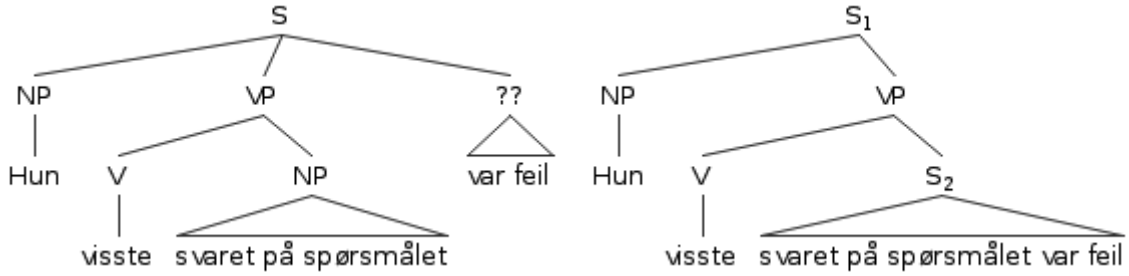
2.3.2 Symmetri

Miller og Chomsky (1963, s. 421-422) diskuterer modeller av språkbrukere og har som utgangspunkt at teoretisk viktige aspekter ved språkevne må være felles for produktive og reseptive funksjoner. Ut fra et syn på språkevne som medfødt virker dette som et fornuftig standpunkt; en medfødt språkevne vil mest sannsynlig ha måttet utvikle seg fylogenetisk gjennom et samspill mellom utvikling av produktive og reseptive evner.

At modellen eller prosesseringsenheten er identisk, betyr imidlertid ikke at prosesseringen trenger å være symmetrisk. Akkurat som at løsning av tredjegradsligninger er mer krevende enn utregning av verdien av et tredjegradsuttrykk, *kan* det være slik at resepsjon er en mer komputasjonelt krevende prosess enn produksjon. Det er også mulig at enkelte konstruksjoner, for eksempel lokal tvetydighet, dobbelt negering eller visse typer av rekursjon, gjør reseptiv prosessering mer krevende enn produksjon, mens det kan være motsatt i andre tilfeller.

Et typisk eksempel på hva som trolig representerer asymmetrisk prosesseringskompleksitet, er såkalte *garden path*-setninger, altså setninger med lokal tvetydighet, der mottakers parsing ofte resulterer i midlertid feilanalyse pga. at det undervegs dannes en hypotese om en annen trestruktur enn den endelige. Setningen i (17) er et eksempel.

(17) Hun visste svaret på spørsmålet var feil.



Treet til venstre representerer en midlertidig feilparsing, mens treet til høyre viser den korrekte strukturen.

Frazier (1985, s. 135-) peker på at slik feilparsing øker kompleksiteten i et resepsjonsperspektiv, men det er lite trolig at *garden path*-konstruksjoner medfører økt kompleksitet i produksjonen. I hvert fall må en hypotese om det påvises eksperimentelt. Men intuitivt virker det ikke slik. Vi vet at vi til stadighet produserer setninger der preposisjonsfraser har uklar plassering i setningsstrukturen, og at det heller kan være krevende å sørge for å uttrykke de samme forholdene utvetydig.

Også andre typer konstruksjoner, for eksempel senterinnføying, kan tenkes å ha asymmetrisk komputasjonell kompleksitet, både teoriorientert og brukerorientert. I kritikk av Yngves dybdehypotese (f.eks. Frazier, 1985, s. 154) er det kommet fram at mange språk har mye sterkere tendens til venstrevridde trær enn hva som er vanlig i norsk og engelsk, og selv om Yngve senere (1998, s. 633-635; 2006, s. 9) har tilbakevist den vitenskapelige verdien av sin

egen hypotese ved å påvise at den ikke er testbar, tyder dette på at eventuelle asymmetriforhold i språkprosessering også kan være språkavhengig.

Det er også sannsynlig at slike eventuelle forskjeller avhenger av prosesseringsenheten, og at symmetriforhold dermed kan variere fra individ til individ, jf Chiperes eksperimenter med individers ulike evner til å prosessere komplekse syntaktiske strukturer (Chipere, 2009). Chipere refererer til tidligere arbeider som avdekker slike forskjeller, men gjengir også sine egne resultater som viser at universitetsutdannede *uten* engelsk som morsmål skårer bedre enn universitetsutdannede personer *med* engelsk som morsmål, noe som skulle tyde på at grammatisk kompetanse kan utvikles ved øvelse eller instruksjon, og at morsmålskompetanse dermed ikke er "fullstendig".

2.3.3 Redundans

Redundans av McWhorters type øker som nevnt kompleksiteten i systemet, men det er også trolig slik at redundans øker produksjonskompleksiteten, i form av antall regler som må utføres for å konstruere en ytring. Dette er altså et tilfelle av regelkompleksitet. Dette vil også øke kompleksiteten i beskrivelsen av trestrukturen til en ytring, ettersom trestrukturen vil inneholde mer informasjon.

I et komputasjonelt perspektiv kan redundansen imidlertid ofte likevel bidra til å lette resepsjonen, både generelt ved å tilføre informasjonsteoretisk redundans i signalet og slik sett øke toleransen for ulike typer støy, og spesielt ved å bidra til å avklare den semantiske strukturen i ytringen, som illustrert i eksemplet i (18).

(18) Hun eide et rødt hus_i, men da hun kjøpte ei hytte, malte hun det_i $gult_i$. [BUJ]

Imidlertid er den type obligatorisk redundans som McWhorter diskuterer, ikke så interessant i forbindelse med problemstillingen i denne avhandlingen, ettersom det ikke er slik at ytreren kan velge å tilføre denne redundansen eller ikke. Utspringet er i systemet, ikke i bruken.

I et ytringsperspektiv er det imidlertid slik at ytreren også kan velge å bruke flere ord eller mer omfattende konstruksjoner enn hva som er nødvendig for å overbringe betydningen. Slik sett tilfører ytreren redundans til signalet – altså ekstra kompleksitet, noe som kan lette resepsjonsprosesseringen. *Garden path*-setningen i (17) kan gjøres mindre resepsjonskompleks ved å bruke en subjunksjon i subklaususen, og dette vil kunne være et mer eller mindre bevisst valg fra ytrerens side:

(19) Hun visste at svaret på spørsmålet var feil.

Andre studier viser at tilføring av redundans av akkurat denne typen, er et trekk som er vanligere i planlagt, prototypisk skriftlig språk (f.eks. Biber, 1988, s. 89).

2.3.4 Tekstkompleksitet

Et trivielt utgangspunkt for å snakke om tekstkompleksitet er gjennomsnittlig strukturell kompleksitet eller andre former for gjennomsnittlig setningskompleksitet. Slik er alle

perspektivene på strukturell kompleksitet også relevante for tekst som helhet. En tekst med større gjennomsnittlig dybde kan sies å være mer kompleks enn en tekst der setningene gjennomsnittlig er mindre komplekse.

2.3.4.1 Syntaktisk variasjon

Men et mer interessant perspektiv på tekstkompleksitet er variasjon, som kan måles blant annet ved hjelp av entropi. For eksempel bruker Gries (2009, s. 112) entropi til å vurdere kompleksiteten i substantivfraser i en tekst. I en tekst på engelsk med 300 substantivfraser sammenligner han substantivfraser med ubestemt artikkel, med bestemt artikkel og uten determinativ. Da vil 100 fraser av hver type representere maksimal uorden og ha effektivitet E = 1, mens et ekstremt tilfelle med 300 fraser av én av typene vil ha effektivitet E = 0 og maksimal "orden" i informasjonsteoretisk forstand:

(20) $E(100, 100, 100) = -3 * (⅓ * \log(⅓)) / \log(3) = 1$

(21) $E(0, 0, 300) = -(2 * (0 * \log(0)) + (1 * \log(1))) / \log(3) = 0$

En spredning av for eksempel subkategorier av subklaususer kan også relateres til en teksts kompleksitet; en tekst med nesten bare nominalklaususer vil være mer ordnet og ha lavere entropi enn en tekst der subklaususene er likt fordelt mellom nominale, relative og adverbiale. Tilsvarende vil en tekst der klaususene eller t-enhetene har stor spredning i lengde, være mindre ordnet og ha større kompleksitet enn en tekst der alle segmentene av samme type er like lange.

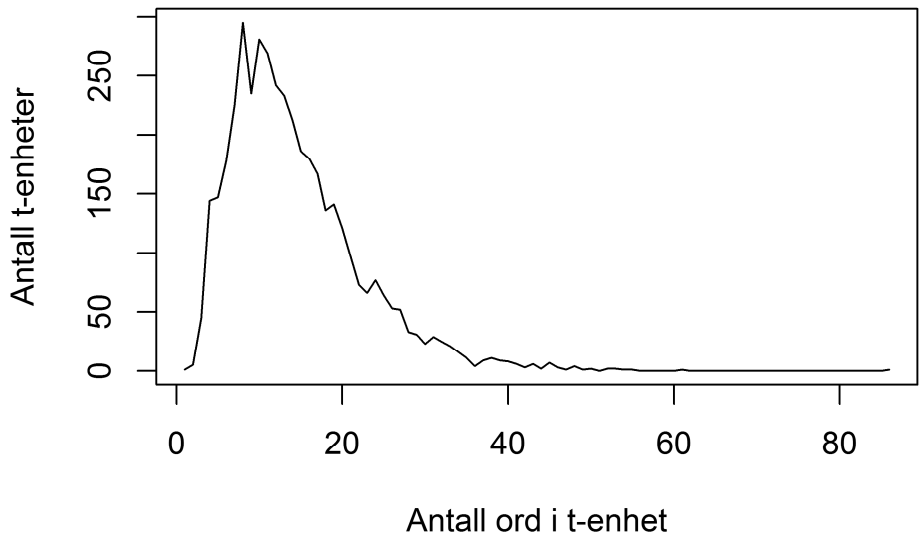
Informasjonsteoretisk er nok dette ikke helt relevant, for i en kommunikasjonssituasjon vil det sjelden være aktuelt å overføre informasjon om segmentlengde eksplisitt. Men entropi kan likevel brukes til å måle variasjonen eller kompleksiteten i mengden, og disse uttrykkene for tekstlig kompleksitet stemmer godt overens med et intuitivt inntrykk av en teksts kompleksitet eller *modenhet*. I et brukerrelatert perspektiv er dette først og fremst et uttrykk for kompleksitet i *produksjon*; det er mer krevende å produsere en variert tekst. Selv om vi i resepsjon kan oppfatte den økte variasjonen, resulterer den ikke i at resepsjonsprosessen blir mer krevende.

Imidlertid er det problematisk å regne ut entropi ut fra en råskåre på denne måten. Uorden målt i entropi forutsetter at en uniform fordeling er den minst ordnede, altså den distribusjonen vi har ved kast med én terning, der alle verdiene har lik sannsynlighet.

For fordelingen av de tre subkategorier av subklaususer stemmer dette ganske godt, selv om frekvensen av hver subkategori neppe er helt lik i et stort tekstkorpus. I elevkorpuset er fordelingen ca. 5 : 4 : 3 mellom nominale, relative og adverbiale klaususer; de eksakte tallene er 1789 : 1439 : 987. Ved vurderingen av kompleksitet i en enkelt tekst kunne man tenke seg å vekte frekvensene mot et slikt standardmål, slik at entropisk effektivitet = 1 tilsvarer en tekst med samme mål som standardmålet, og at et avvik fra dette ville gi lavere entropi. Dette ville være en form for relativ kompleksitet. Det er imidlertid ikke trivielt hvordan en økning i den minst frekvente variabelen skulle påvirke "relativ entropi" i en slik

vekting; informasjonsteoretisk vil en slik økning medføre tapt komprimeringspotensial, så kanskje må en annen type mål i så fall benyttes.

For variabelen setningslengde – eller i denne studien egentlig t-enhetslengde – støter vi på andre problemer. I elevkorpuset har distribusjonen av variabelen en tydelig pukkel og en hale mot høyre (figur 2-2); det er rimelig å anta at distribusjonen av t-enhetslengde følger et lignende mønster i de fleste tekstkorpus, muligens prinsipielt en lognormal fordeling (se 7.2.2.5 om lognormale distribusjoner).



Figur 2-2: Fordeling i elevtekstkorpuset av t-enheter etter lengde

Intuitivt virker entropimålet som et godt mål på kompleksitet i den forstand at en tekst med høyere spredning i t-enhetslengde får høyere entropi enn en tekst med lavere spredning i t-enhetslengde. Imidlertid kan det hende vi ville reagere på en tekst med tilnærmet uniform fordeling av t-enhetslengder og kanskje ikke kalle den *moden*. Men vi ville kanskje kalle den *kompleks,* vanskelig eller merkelig – for kompleks for vår smak. Dette gjør entropi til et mindre aktuelt mål for denne type tekstkompleksitet.

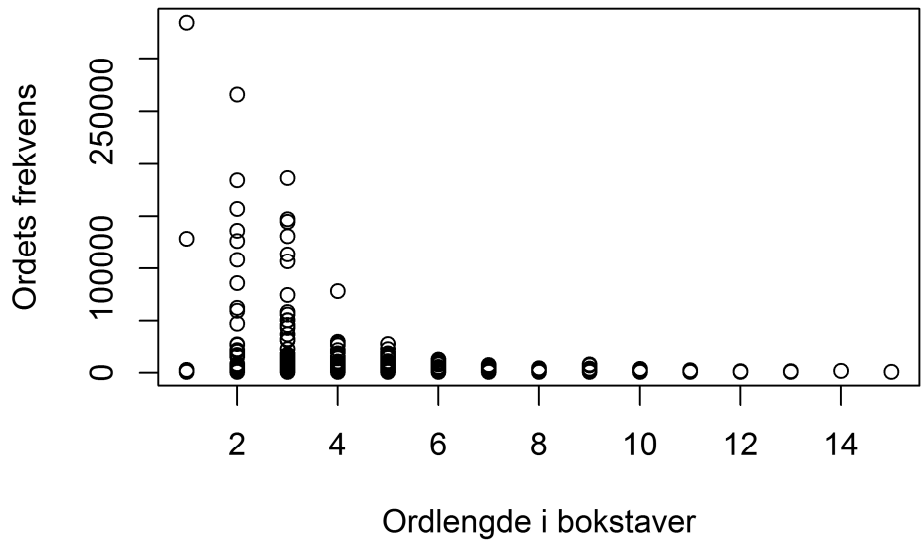
Det er for øvrig også tekniske problemer knyttet til å regne ut entropien i denne fordelingen. Her er nemlig ikke antall ulike størrelser kjent. Selv om den lengste t-enheten i elevtekstkorpuset er på 86 ord, er ikke sannsynligheten for forekomsten av en setning på 87 ord lik null. For å kunne beregne entropien må vi kjenne størrelsen på mengden av antall mulige symboler eller meldinger, men denne variabelen er prinsipielt ubundet. Det medfører at det er vanskelig eller kanskje prinsipielt umulig å regne ut sannsynlighetene.

2.3.4.2 Leksikalsk variasjon

For leksikalsk distribusjon vil høy entropi være et uttrykk for at høyfrekvente ord blir brukt mindre enn i andre tekster, og enten at lavfrekvente ord har høyere frekvens, eller at flere lavfrekvente ord blir brukt, altså at den leksikalske diversiteten er høyere. Begge deler stemmer godt med en intuitiv oppfatning av kompleksitet på leksikalsk nivå. Ekstreme

fordelinger av nærmest uniform karakter – som ved terningkast – er vanskelig å tenke seg, blant annet fordi fordelingen mellom funksjonsord og leksikalske ord vanskelig vil kunne bli lik i noen tekst som er lengre enn en viss nedre grense. Dette betyr trolig at entropimål av leksikalske distribusjoner først og fremst er interessant innenfor én og én ordklasse. Leksikalsk statistikk er imidlertid svært komplisert, blant annet fordi det er mange enheter med ekstremt lave sannsynligheter, en problemstilling som gjerne går under betegnelsen LNRE – *Large number of rare events*. Dette medfører at det er prinsipielt umulig å regne ut sannsynligheten nøyaktig for noen enhet. (Baayen, 2001, s. 229)

Det er mange som har forsøkt å konstruere statistiske modeller for leksikalsk distribusjon, for eksempel Zipfs lov, som sier at et ords rangering i frekvens er omvendt proporsjonalt med dets frekvens (Baayen, 2001, s. 15). Det er dermed mulig å beregne hvorvidt slike distribusjoner er ulike i ulike tekstutvalg, for eksempel ved kurvaturen og stigningstallet i tekstenes Zipf-kurver, slik jeg illustrerer i figur 10-3. Slike sammenligninger kan imidlertid være komplekse og må tolkes for å kunne si noe om forskjeller i tekstlige egenskaper, mens entropi faktisk plasserer en distribusjon på en skala fra orden til uorden, og graden av orden er tett koblet til kompleksitet i form av Kolmogorovs kompleksitetsbegrep. På en måte er leksikon ferdig komprimert, ettersom det er en negativ sammenheng mellom et ords lengde og dets frekvens, slik det går fram av figur 2-3 nedenfor. Spearmans korrelasjonskoeffisient rho ≈ –0,35 blant de 10 000 mest frekvente ordformer i Oslo-korpuset (Tekstlaboratoriet, 2010). Frekvente ord har altså korte koder, nøyaktig slik en LZ-algoritme ville resultere i.



Figur 2-3: Forholdet mellom et ords lengde og dets frekvens, basert på Oslo-korpuset.

2.4 Komputasjonell kompleksitet

Dahl (2009, s. 51-52) peker på at lingvistisk kompleksitet ofte blir forbundet med prosessering, blant annet hvor mye som kreves for å produsere og forstå ytringer på et språk. Det er dette jeg i mitt begrepsapparat vil kalle brukerrelatert komputasjonell kompleksitet, men Dahl ønsker å la slike brukerrelaterte forhold stå utenfor det vi forstår ved lingvistisk

kompleksitet, og heller benevne dette med ord som "*cost*", "*difficulty*" eller "*demandingness*".

2.4.1 Kompleksitet og kostnad

I dette prosjektet er imidlertid brukerrelaterte momenter sterkt medvirkende. Prosjektets hovedhypotese er at fysiske – tidsmessige – rammebetingelser ved ytringssituasjonen påvirker ytringens kompleksitet, og at den høyere produksjonshastigheten ved tastede ytringer medvirker til dette. I dette perspektivet virker det unaturlig å ikke ta komputasjonell kompleksitet med i betraktning i behandlingen av en ytrings kompleksitet. Videre virker det unaturlig å diskutere komputasjonell kompleksitet uten å ta med i betraktning det faktiske maskineriet som normalt brukes til å produsere og resipere språk, nemlig hjernen.

Chomsky-hierarkiet forsøker å skille mellom språk av ulik systemisk kompleksitet ved å legge ulike begrensninger på beskrivelsene av dem. Men hierarkiet er tett knyttet til nettopp strukturell og komputasjonell kompleksitet, ettersom det også definerer hva slags type maskineri som er nødvendig for å generere og parse språkets setninger. Dette maskineriet er matematisk – og ikke brukerrelatert – definert, og forholdet mellom grammatikk og maskinerikompleksitet er entydig begge veier. En type-3-grammatikk krever en finitt automat (*finite automaton*), mens en type-2-grammatikk krever en stakkautomat (*push-down automaton,* PDA) og en type-0-grammatikk krever en *Turing machine*. (Beckman, 1980, s. 331-333).

Disse sammenhengene er interessante for informatikere fordi de sier noe om komputasjonell kompleksitet, relatert til hvilken tid man kan regne med at det tar å parse en "ytring" i et programmeringsspråk, altså et dataprogram, men for lingvister er de mest interessante i den grad de kan fortelle oss noe om produksjon og resepsjon av naturlige språk. Hvis for eksempel naturlige språk kan sies å være kontekstfrie men ikke regulære, er det naturlig å spørre seg om faktisk språkprosessering i hjernen er stakkbasert, og om det er noen begrensninger på denne stakken. I den abstrakte automatteorien er selvfølgelig stakken infinitt, men i den menneskelige hjerne, som er av endelig størrelse, kan det ikke finnes hukommelsesenheter med uendelig kapasitet. I så fall er de heller ikke stakker i automatteoretisk forstand, men de kan i praksis likevel fungere som stakker.

2.4.2 Asymmetri

Forholdet mellom kompleksitet og kostnad henger delvis sammen med asymmetri i komputasjonell kompleksitet – eller prosesseringskompleksitet – mellom produksjon og resepsjon. Mitt eksperiment er først og fremst knyttet opp mot produksjonssituasjonen og rammebetingelser for den, og resepsjonsprosesseringen er bare indirekte relevant ved at skriveren undervegs vil lese det han/hun har produsert, og evaluere teksten med tanke på hva som er vanskelig å resipere. Dessuten er resepsjonssituasjonen indirekte til stede gjennom at skrivere med utviklet mottakerbevissthet vil ha resepsjonen i tankene under produksjonen og eventuelt justere adferden i retning av hva som oppfattes som en mer leservennlig tekst.

Denne dualismen gjør prosesseringsmekanismene i slike tilfeller svært vanskelige å studere. Man studerer prosesseringskompleksitet i produksjonen, mens ytreren – i hvert fall de av dem som har tatt til seg noe av skriveopplæringen – er opptatt av å senke kompleksiteten i *resepsjon*; for ytreren – og for teksten – er i realiteten produksjonskompleksiteten irrelevant. I vår leser-rettede tekstkultur skal teksten se "lett" ut, og den ideelle tekst etterlater seg ikke spor av hvor vanskelig den var å produsere.

2.4.3 Nevrolingvistiske studier

Kunnskap om den menneskelige språkevnen har i de største deler av menneskehetens historie vært tilgjengelig for studier bare indirekte, gjennom ulike former for psykolingvistiske studier eller eksperimenter. Men med de fremskritt hjerneforskning generelt og nevrolingvistikken spesielt har gjort de siste årene, vet vi nå også mer om hva som faktisk skjer i hjernen vår ved språkprosessering. Vi står ved terskelen til å forstå mer om hvorvidt kontekstfrie grammatikker og stakkautomater bare er abstrakte størrelser, eller om de har fysiske paralleller i hjernen. Dette vil bringe oss mer kunnskap rundt i hvilken grad de abstrakte betraktningene rundt språkprosessering, inkludert praktiske datalingvistiske arbeider, har fysisk realitet i hjernen. For eksempel kan vi kanskje få svar på spørsmålet om hvorvidt klausal subordinering faktisk forutsetter en type prosessering i hjernen som kan forsvare å hevde at denne typen subordinering øker en ytrings komputasjonelle kompleksitet.

Friederici & Brauer (2009) (referert i Givón & Shibayama (2009)) påviser faktisk at parsing av type-2-strukturer av typen A^nB^n aktiverer områder av hjernen som er passive under parsing av type-3-strukturer av typen $(AB)^n$. Dette tyder i det minste på at parsing av visse typer av klaususunderordning forutsetter en annen type maskineri, og hvis dette er en forutsetning, er det mye som tyder på at dette kan beskrives som økt komputasjonell kompleksitet – ikke bare objektivt sett, men også relatert til dette spesifikke parsingsmaskineriet.

2.5 Relevans for en studie om skriveverktøy

2.5.1 Det brukerrelaterte perspektivet

Brukerrelatert kompleksitet er avhengig av perspektivet; de samme konstruksjonene er ikke vanskelige for taler eller lytter, og for L1- eller L2-tilegnere (Miestamo, 2008, s. 25-26). Både i dette prosjektet og generelt må vi dessuten legge til aktørene *skriver* og *leser*. Hva som er vanskelig for taler, er ikke nødvendigvis det samme som er vanskelig for skriver, og tilsvarende for lytter/leser, jamfør for eksempel funn som tyder på at talte og skrevne tekster har ulike grenser for maksimal kompleksitet i klausal underordning (Karlsson, 2009, s. 193).

Dette forholdet kompliseres ytterligere av at jeg sammenligner skriving med to ulike skriveverktøy og hypotetiserer at skriveverktøyet kan påvirke ytringskompleksiteten, uten at jeg vet mye om hva slags faktorer i en ytringssituasjon som kan tenkes å påvirke kompleksiteten.

Det er også uklart i hvilken grad det er mulig å konkludere om komputasjonell kompleksitet i produksjonen gjennom å studere den strukturelle eller statiske kompleksiteten i produktet, slik jeg gjør i prosjektet.

2.5.2 Testing av teorier

Jeg har i dette kapitlet omtalt flere ulike teorier om kompleksitet generelt og om klausal dybde spesielt. Det ville være interessant å gjøre analyser av det korpusmaterialet som er samlet inn i prosjektet, basert på hver enkelt teori og se om noen av dem gav positive forskjeller mellom håndtekster og tastaturtekster og andre ikke. Dette ville i så fall være en støtte til den aktuelle teorien, riktignok en svak støtte.

Et eksempel er en test av Karlssons hypotese om maksimal klausal dybde på 5 i skriftlig språk (Karlsson, 2009, s. 200). T-enheten i (22) er den t-enheten i korpuset som har den dypeste klausale innføyingen, nemlig ned til nivå 6.

(22) {T Og til slutt har vi det svaret

{S som jeg tror
{S redaktøren i Småby Arbeiderblad mente
{S da denne kom med påstanden om
{S at svaret bør være enkelt
{S når ungdommer spør foreldrene sine
{S om de kan få med øl til en fest;}}}}}}}

Den overskrider dermed Karlssons "stilistiske preferanse" F^5max-w og skulle fremstå som kompleks for en mottaker. Intuitivt oppfatter jeg ikke setningen som særlig kompleks eller vanskelig å forstå, og jeg tror heller ikke skriveren gjør det.

Det er generelt et problem for en rent korpuslingvistisk metode som den jeg benytter, at man ikke har tilgang på hverken skriveprosessen, skriverens vurdering eller intensjon eller lesingen eller leserens oppfatning. Psykolingvistiske og nevrolingvistiske eksperimenter ville kunne utfylle den korpuslingvistiske metoden i slike sammenhenger. En analyse av skriveprosessen ville for eksempel kunne avdekke om en slik t-enhet ble produsert lineært som i en taleytring, eller om produktet er et resultat av omfattende redigering og kanskje som en følge av redigeringen til og med å oppfatte som en performansefeil.

Et mer teoretisk spørsmål knyttet til eksemplets kompleksitet er hvorvidt den klausale underordningen i (22) kan sies å være rekursiv i det hele tatt. I et komputasjonelt perspektiv er det kanskje heller rimelig å anta at final underordning av den typen som alle subklaususene i eksemplet representerer, representerer iterative konstruksjoner som ikke fordrer mer enn iterativ prosessering, noe som ikke krever et stakkbasert apparat i det hele tatt, men kan prosesseres i type-3-apparatur.

2.6 Oppsummering

Tittelen på dette kapitlet er "Språklig kompleksitet". Det skulle nå være klart at "språklig kompleksitet" neppe kan sies å være ett begrep, men en betegnelse som brukes om flere mer eller mindre nært beslektede begreper, der hovedskillelinjen må sies å gå mellom kompleksitet i språksystemer og kompleksitet i språkytringer. Noen aspekter har paralleller i disse to hovedtypene av språklig kompleksitet, mens andre er inkompatible. Når det gjelder kompleksitet i ytringer, er det et prinsipielt skille mellom strukturell kompleksitet i t-enheter og variasjon i tekster. Ifølge Halliday finnes det også (minst) to ulike dimensjoner av strukturell kompleksitet, knyttet til henholdsvis innfløkthet og informasjonstetthet eller kompakthet.

Et annet viktig skille går mellom teoriorientert kompleksitet og språkbrukerrelatert kompleksitet. Jeg tror jeg har greid å vise at noen matematiske og naturvitenskaplige begreper har anvendelse i lingvistikken, og jeg tror også at resultater fra naturvitenskaplige eksperimenter som nevrolingvistikk i fremtiden vil kunne belyse relativ språklig kompleksitet på viktige måter.

3 Skriving og registervariasjon

Mens forrige kapittel hadde fokus på språklig kompleksitet, bygger dette kapitlet et teoretisk fundament knyttet til registervariasjon og modus og til skriving generelt og digital skriving spesielt.

3.1 Modus, stil og språkvariasjon

Prosjektet i denne avhandlingen bygger på forskning på register, og spesielt på forskjeller mellom skrift og tale. Et tradisjonelt syn på skrift og tale er at de befinner seg i hver sin ende av en kompleksitetsskala med flere trekk som typisk opptrer sammen. Ifølge det tradisjonelle synet er enklere trekk som parataksе, setningsfragmenter og konkreter typiske muntlige trekk, mens mer komplekse trekk som hypotakse, fullstendige setninger og abstrakter er typisk for skrift. Et eksempel på en slik tradisjonell presentasjon finnes i Eiliv Vinjes *Tekst og tolkning*:

> Karakteristisk for munnleg språkbruk er gjentakingar, bruk av sentensar, klisjear, ordspråk, "ufullstendig" setningsstruktur og sidestilt (parataktisk) samanføying av setningar. Det talte språket er adderande, føyer til ny informasjon litt etter litt. Skriftspråket er analytisk, syntetiserande og systematiserande. Jamført med talen er det langt meir hypotaktisk både i setningsstruktur og i samanføyinga av setningar, og ikkje minst er setningane "fullstendige" (E. Vinje, 1993, s. 46).

Noen talesjangre, for eksempel en forberedt forelesning, vil være mer skriftlige i formen, mens noen skriftlige sjangre, for eksempel personlige brev, kan være mer muntlige.

Ifølge Michael Halliday (1989, s. 86; 1998) er det ikke riktig at muntlige ytringer typisk er parataktiske; muntlige ytringer består tvert imot av varierte, innfløkte konstruksjoner (1989, s. 76-) der setninger kombineres gjennom forskjellige typer mekanismer. Halliday mener at det ikke er riktig at muntlig er mindre komplekst enn skriftlig, men at det derimot er slik at muntlig og skriftlig språk er komplekst på ulike måter. Planlagte, skriftlige tekster er gjerne mer kompakte (1989, s. 62); de har altså mer informasjon per enhet og større leksikalsk tetthet (1989, s. 61), gjerne ved at antallet grammatiske ord er større, mens antall leksikalske ord er omtrent det samme (1989, s. 80). Muntlige tekster er derimot mer syntaktisk innfløkte, med større grad av hypotakse og lavere gjennomsnittlig ordformlengde. Kompleksiteten i tale er grammatisk, mens kompleksiteten i skrift er leksikalsk. Han argumenterer for dette i flere arbeider (Halliday, 1979, 1987, 1989), men de illustrerende eksemplene er konstruerte, og han viser ikke til noen autentisk empiri som underbygger påstandene.

Wallace Chafe viste derimot i flere empirisk baserte arbeider, delvis sammen med medforfattere, (Chafe, 1982; Chafe & Danielewicz, 1987; Chafe & Tannen, 1987) at en endimensjonal modell av språklige trekk er en for enkel tilnærming til analyse av muntlig og skriftlig språk. Han påviste to uavhengige dimensjoner, og knyttet dem til to ulike situasjonelle faktorer, nemlig ulike begrensninger i tidsrammene for kognitiv prosessering på den ene side, og forholdet til et "publikum" (Chafe, 1982, s. 52) eller mer generelt

kommunikativ interaksjon på den andre. Chafe og Danielewicz finner høyere TTR (se kapittel 10) i skriftlig kommunikasjon enn i muntlig (1987, s. 88), mer sammentrekninger (s. 94) og kortere *intonation units* (s. 96). Men når det gjelder preposisjonsfraser (s. 98) og nominaliseringer (s. 100) og attributive adjektiver (s. 101), er ikke bildet like enkelt; de finner færrest preposisjonsfraser, færrest nominaliseringer og færrest attributive adjektiver i samtaler, flere i forelesninger og i brev, og flest i skriftlige akademiske tekster. At distribusjonene ikke følger skillet mellom muntlig og skriftlig språk, viser at det finnes mer enn én type av påvirkningsfaktorer. Chafe og Danielewicz undersøker flere variabler og finner flere ulike mønstre fordelt etter de fire teksttypene de undersøker, men disse eksemplene er tilstrekkelig til å underbygge konklusjonen om at variasjonen over muntlig og skriftlig språkbruk ikke er endimensjonal. Forfatterne tilskriver de mønstrene de finner, at trekkene påvirkes av både kognitive og kontekstuelle faktorer, og at disse to typene av påvirkning er prinsipielt uavhengige av hverandre.

I tradisjonell skriveteori er prosesseringsrammebetingelser regnet som en del av skrivesituasjonen, slik for eksempel Rijlaarsdam og van den Bergh (2006) gjør i sine forsøk, men Chafes skille innebærer at aspekter som har med fysiske rammebetingelser å gjøre, for eksempel tid, isoleres fra betingelser som er mer relatert til personlig interaksjon, mange av dem knyttet til Jakobsons språkfunksjoner (for eksempel Vagle, Sandvik, & Svennevig, 1994). Vagle (1990, s. 123-124) er inne på det samme når hun sier at tidligere tale/skrift-forskning blander "sammen genre/kommunikativ hensikt og modalitet".

Douglas Biber (1986, 1988) tar utgangspunkt i den samme antagelsen som Chafe, nemlig at språkbruksvariasjon henger sammen med flere uavhengige påvirkningsfaktorer, og at trekkene derfor må variere over flere dimensjoner enn én. Han refererer til Chafe (1982) når han deler rammebetingelsene for ytringer opp i prosesseringsaspekter og situasjonsaspekter, men han har en mer avansert tilnærming, der han samler inn et stort antall tekster i et stort antall ulike teksttyper og analyserer et stort antall variabler med en eksplorerende multivariat analysemetode kalt faktoranalyse.[3] Biber bekrefter at en endimensjonal modell av språklige trekk er for enkel, og at språk varierer over flere uavhengige dimensjoner, der muntlighet og skriftlighet ikke er den mest fremtredende faktoren. Bibers (1986) mer nyanserte syn er dermed at språklige trekk i tekst vil variere med funksjoner som interaktivitet kontra redigeringsgrad, abstrakt kontra situert innhold, og rapportert kontra umiddelbar stil, kanskje ved siden av modus. Det som tradisjonelt har vært omtalt som muntlige eller skriftlige trekk, kan gjerne tilskrives enkelte av disse tre dimensjonene, men ikke nødvendigvis alle. I 1988 (s. 104-) videreutvikler og delvis justerer Biber de dimensjonene han kom fram til i 1986, men den mest fremtredende dimensjonen av interaktivitet kontra redigeringsgrad består i hans nye analyse med 6 dimensjoner. Et interessant resultat er at det som var Chafes utgangspunkt, nemlig at interaksjon og stramme kognitive rammer er uavhengige dimensjoner, blir slått sammen i den viktigste dimensjonen i Bibers analyser, og som Biber

[3] Faktoranalyse er beslektet med prinsipalkomponentanalyse, som jeg forklarer i 12.1.

selv peker på, er dette egentlig ikke så rart, ettersom dialogisk interaksjon nettopp legger strenge rammer på produksjonen, mens monologisk produksjon gir god mulighet for planlegging og redigering.

Det jeg prøver å gjøre i dette prosjektet, er å isolere en viss type prosesseringsbegrensninger for å undersøke i hvilken grad disse påvirker språklige trekk i skriftlige tekster. Av Bibers tre eller seks dimensjoner (1986; 1988) er det særlig dimensjonen med ytterpunktene interaktiv kontra redigert tekst som er interessant for prosjektet, ettersom de andre dimensjonene er tettere knyttet til teksttype og språkfunksjon enn til modus. Blant de språklige variablene som er fremtredende i Bibers første dimensjon, er de *interaktive* trekkene *that*-stryking, pronomenet *it*, *be* som hovedverb, kausative subklaususer, relativklaususer, *WH*-klaususer og finale preposisjoner, og de *redigerte* trekkene substantiver, ordlengde, preposisjoner og attributive adjektiver (1988, s. 102)[4]. Det er påfallende at flere av disse trekkene er direkte refleksjoner av Hallidays ikke-empiriske skille mellom innfløkthet og kompakthet. Det er også verdt å merke seg at Biber i analysen skiller mellom ulike typer av subklaususer, og at de har ulike egenskaper i de ulike dimensjonene. Relatert til dette er Hallidays skille mellom hypotakse og innføying, som han karakteriserer som henholdsvis iterasjon og rekursjon og tilskriver helt ulike kognitive egenskaper (1989, s. 82-84). Dette er et skille jeg ikke har greid å utnytte i denne studien.

Collot og Belmore (1996) peker på at de konkrete variablene i en slik analyse vil variere med språk, og rapporterer at Saukkonen (1989, 1993) for eksempel opererer med andre variabler for finsk. Variabler som sammentrekninger (*it's*) og hjelpeverbet *do* er typiske eksempler på variabler som er språkspesifikke og ikke kan overføres til norsk. Derimot er variabler som ulike typer av subjunksjoner (*that, if, wh-*), pronomenet *it*, splittede preposisjonsfraser, ordlengde og type/eksemplar-forhold (*type/token ratio*, ofte forkortet til TTR, også på norsk) variabler som godt lar seg overføre til norsk. Om de alle har samme relative rolle i et flerdimensjonalt rom av språkbruksvariabler, er noe mer usikkert, men det virker ganske trygt å anta at det finnes paralleller når det gjelder for eksempel sammenhengen mellom gjennomsnittlig ordlengde og TTR og hvor kognitivt krevende produksjonen er.

Baron (1998) refererer til Bibers flerdimensjonale analyse (1988), men velger likevel å arbeide innenfor en tale/skrift-dikotomi i sin analyse av språket i epost. Hun bruker muntlige kjennetegn som lavt type/eksemplar-forhold, lav leksikalsk tetthet, færre adverbiale leddsetninger og færre disjunksjoner (\eller\).

3.2 Skriveprosess og skriveferdigheter

Skriving er en kompleks prosess som er satt sammen av mange ulike typer delprosesser, som gjerne er inndelt i hovedgruppene planlegging, overføring og gjennomsyn. Prosessorientert

[4] Alle Bibers variabler bortsett fra ordlengde og TTR er regnet ut som antall forekomster per 1000 ord.

skriveteori ser delprosessene ikke som lineære, men som rekursive og potensielt samtidige: "Disse prosessene har en hierarkisk og sammenvevd organisering, der en gitt prosess kan ligge innbygd i en hvilken som helst annen prosess." (Flower & Hayes, 1991, s. 103) Omskriving kan for eksempel foregå mens skriveren holder på å skrive førsteutkast. Å være en god skriver forutsetter derfor at man har gode ferdigheter av mange forskjellige typer, og det forutsetter gode metakognitive evner – at man er flink til å overvåke prosessene og disponere ressursene. Også korttidsminne og langtidsminne spiller en rolle her, og bakgrunnskunnskaper om emne, mottakere, teksttyper og skrivestrategier er viktige. Svake skrivere kan derfor være svake av helt ulike årsaker.

Skriving med tekstbehandlingsprogram krever visse andre ferdigheter til erstatning for eller i tillegg til allmenne skriveferdigheter. Skriving på tastatur stiller trolig lavere finmotoriske krav til skriveren, men det forutsetter uansett trening å komme opp på et akseptabelt hastighetsnivå. Dessuten krever det ganske stor grad av automatisering å kunne se på produktet og ikke på den motoriske aktiviteten mens man skriver; i håndskrift kan man se på begge deler samtidig. Russell (1999) testet tastehastighet hos sine forsøkspersoner ved avskrift av to tekster over to minutter og regnet ut antall ord per minutt. Gjennomsnittsverdien for hans utvalg av åttendeklassinger (*grade eight*) var 17 ord per minutt, med 5 og 38 som høyeste og laveste verdi. Han skriver at et vanlig krav for ansettelse av sekretærer er 40, og at hans forsøkspersoner dermed stort sett må ansees for å være begynnere i tastaturskriving. Han rapporterer også at de fleste av forsøkspersonene ser på tastene mens de skriver – også de raskeste skriverne. I en annen undersøkelse av forskjeller mellom skriving for hånd og med maskin hos voksne (Horowitz & Berkowitz, 1964) ble forsøkspersoner som produserte færre enn 45 ord per minutt på skrivemaskin, ikke regnet med. Forholdet mellom ord per minutt og tastetrykk per minutt vil selvfølgelig variere med både teksttype og språk. Dette er derfor ikke et nøytralt mål. Det er også sannsynlig at hastigheten vil påvirkes av tekstens vanskelighetsgrad, av skriverens kjennskap til emnet, og av andre variabler som påvirker lesing (L. I. Kulbrandstad, 2003); Russell har valgt leksikonartikler for sin test. Horowitz og Berkowitz (1964) bruker en standardtest fra Facit Corporation.

I tillegg til det rent mekaniske kreves det kjennskap til verktøyets redigeringsfunksjoner, og de mest brukte av disse må også være tilstrekkelig automatisert til at de ikke stjeler kognitiv kapasitet fra tekstskapingen. Russell (1999) undersøker korrelasjonen mellom tekstkvalitet og forsøkspersonenes *pc-kyndighet*, og i tillegg til å teste tastaturhastighet bruker han et spørreskjema der forsøkspersonene rapporterer både om hvor lenge de har brukt pc i ulike sammenhenger, hvor ofte de bruker pc til ulike formål og i hvilken grad de foretrekker pc framfor håndskriving til ulike oppgaver. Resultatene fra Russells undersøkelser er presentert på side 46.

De visuelle rammene for skrivingen spiller også en rolle; det krever andre kognitive evner og en annen bruk av korttidsminnet å arbeide med strukturen i en skjermtekst, som er mindre håndgripelig enn en papirtekst. Flere forskere har arbeidet med hypoteser om at korttidsminnet/arbeidsminnet består av to eller tre ulike komponenter: en verbal/fonologisk

komponent og en eller to komponenter som har med det visuelle og med rom og struktur å gjøre (Kellogg, 2004; Torrance & Galbraith, 2006), og Piolat (1991, s. 260) mener at "small sized sections of text as they appear on the screen [lead writers] to modify mainly the surface aspects of the text." I et senere arbeid (Piolat, Roussey, & Thunin, 1997) viser hun imidlertid at måten teksten presenteres på skjermen på, kan redusere denne effekten.

3.3 Skriveverktøy, skriveprosess og skriveprodukt

Hypotesen i dette prosjektet er knyttet først og fremst til skrivingens hastighet med ulike skriveverktøy, de redigeringsmuligheter som verktøyene gir, og til motivasjon. Dessuten drøfter jeg undervegs muligheten for påvirkning fra språkbruk eller registerkjennetegn i digitale medier. Dette delkapitlet handler om disse fire faktorene.

3.3.1 Hastighet

Forutsatt at tastaturferdighetene er gode nok, er tastaturskriving raskere enn håndskriving. Hastigheten i tekstproduksjonen nærmer seg hastigheten for tanken og taleproduksjon. Dette har med produksjonsrammer for ytringene å gjøre (Biber, 1986; Chafe, 1982), og hvis språkbrukeren utnytter muligheten til å produsere raskt, gjør det rammene mer like den prototypiske muntlige ytringssituasjonen: mindre muligheter for planlegging og redigering før meldingen når mottaker.

En studie fra 60-tallet (Horowitz & Berkowitz, 1964) sammenlignet tekster skrevet med verktøy som gir ulik skrivehastighet: håndskrift, maskinskrift og maskinstenografi. De fant at raskere produksjon resulterte i tekster med lavere TTR, det vil si med et sterkere muntlig preg, dog uten at TTR nådde samme nivå som tale, selv for det raskeste verktøyet, maskinstenografi.

Kellogg (1994) mener imidlertid at produksjonshastighet *ikke* er en viktig ramme for skrivingen, og han rapporterer at profesjonelle skribenter skriver like fort for hånd som med tekstbehandling. Undersøkelsen er imidlertid relativt gammel, og man kan lett tenke seg at dagens ungdommer har et annet forhold til tekstproduksjon enn hva Kelloggs profesjonelle forsøkspersoner hadde for mer enn to tiår siden.

Russell (1999) rapporterer at skribenter med tastehastighet på over 20 ord per minutt og generelt gode pc-ferdigheter skrev lengre og *bedre* tekster på tekstbehandling enn på papir, og resultatet støttes av en senere undersøkelse (Russell & Plati, 2001). Russell undersøkte imidlertid ikke om det er korrelasjon mellom tastehastighet og kvalitet, og undersøkelsen gjaldt heller ikke språklige trekk.

MacArthur (1999) skriver at selv om korrelasjon mellom tekstkvalitet og skrivehastighet er etablert når det gjelder håndskrift, er foreløpig ikke noen slik sammenheng påvist når det gjelder pc-skriving. Han fremhever også bedrede redigeringsmuligheter og motivasjon knyttet til publisering og til produktets estetikk som viktige faktorer i skrivekvalitet og

tekstbehandling. MacArthur skriver riktignok først og fremst om elever med skrivevansker, men om disse tre faktorene skriver han generelt.

3.3.2 Redigering

Redigeringsmuligheter har også med produksjonsrammer å gjøre. Tekster skrevet for hånd kan selvfølgelig også redigeres. Kostnaden er imidlertid større, enten i form av arbeidsinnsats ved at partier av teksten må skrives på nytt, eller ved at presentasjonens estetikk blir skjemmet av overstrykninger eller korrekturlakk. Mens papir-og-blyant-tekster trolig blir endret i et mindre antall diskrete revisjoner, er sannsynligvis tekstbehandlingstekster heller omskrevet i en mer kontinuerlig sekvens av mindre revisjoner. Et slikt endringsmønster kan bidra til at slike tekster i enda større grad enn papirtekster framstår som planlagte og redigerte.

MacArthur (1999, s. 13; 2006) rapporterer at skriveopplæring med tekstbehandling gir bedre tekster, og at denne opplæringen også gjør papir-og-blyant-tekster bedre. Også Bangert-Drowns (1993) rapporterer den samme effekten. Dette indikerer at opplæring i elektronisk tekstbehandling endrer skriveprosessen for disse skriverne, og at denne endringen ikke bare er knyttet til skriveverktøyet i seg selv. Hvilke deler av skriveprosessen som er påvirket, er usikkert, men effekten kan tyde på at skrivere som har lært å nyttiggjøre seg mulighetene i tekstbehandlingsverktøyet, gjør bruk av de samme strategiene når de skriver for hånd.

Kellogg (1994) vurderer forskning på området fram til 1994 og rapporterer at skriveprosessen endres vesentlig med verktøyet, men at man ikke kan finne endringer i tekstenes kvalitet. Andre (Collier & Werier, 1995; Harrington, Shermis, & Rollins, 2000; Hawisher, 1986) rapporterer det samme. Flere forskere (Haas, 1989; Kellogg & Mueller, 1993) rapporterer til og med at tekstene som helhet oftest blir *dårligere* med tekstbehandling. Også flere lærere har i uformelle samtaler med meg antydet at de synes elevenes tekstbehandlede tekster gjerne har dårligere kvalitet enn de håndskrevne. En annen undersøkelse rapporterer at mye tekstbehandlingsredigering skjer i tekstoverflaten (MacArthur, 2006). Det vil si at redigeringen kanskje påvirker det syntaktiske mer enn det tekstlige. Kellogg og Mueller undersøker dessuten kohesjon uten å finne signifikante forskjeller, men ingen av disse studiene analyserer syntaktiske forskjeller.

Fra forskning på prosessorientert skrivepedagogikk kjenner vi også til at sterke elever i større grad enn svake elever klarer å nyttiggjøre seg omskriving som en delprosess i skriveprosessen til å gjøre tekstene bedre gjennom omfattende endringer (Piolat, 1991). Svakere elever redigerer mer i overflaten, altså i det syntaktiske, det morfologiske og det leksikalske (Eritsland, 2004, s. 107-109; Fitzgerald, 1987). Fitzgerald sier videre at majoriteten av forskningsarbeider viser at de fleste endringer skjer i overflaten, og Beach og Friedrich (2006) viser til forskning som viser at over 80% av omskriving (*revisions*) i en 12. klasse var overflateendringer.

De to faktorene hastighet og redigeringsmuligheter kan altså tenkes å virke i motsatte retninger, og kanskje er det slik at redigeringsmulighetene i visse situasjoner mer enn oppveier hastigheten.

3.3.3 Motivasjon

Både forskere og lærere peker på bruk av digitale verktøy som motiverende faktor i skrivearbeid i skolen. I ulike former for reell digital kommunikasjon vil det være sider ved kommunikasjonssituasjonen som virker motiverende, og mer avanserte tekstlige egenskaper som hypertekst og multimodalitet kan virke på samme måte. Motivasjon knyttet til bruk av tekstbehandlingsverktøy alene kan være relatert til aspekter som:

- større skrivehastighet
- bedre redigeringsmuligheter
- mindre slitsomt motorisk arbeid
- penere produkt
- spennende teknologi

Imidlertid kan motivasjonen avhenge av i hvilken grad skriveren er fortrolig med verktøyet, og en del elever rapporterer (f.eks. Russell, 1999) at de arbeider raskere og tryggere med papirskriving, og at de tenker bedre med papir og blyant. Mange skribenter rapporterer at de ikke greier å se på skjermen samtidig med at de skriver.

For mange er nå tastaturskriving normalsituasjonen, og for mange av disse skribentene vil skriving på papir kunne oppfattes som demotiverende, med argumenter som er motpoler til argumentene over: det går sakte, man kan ikke rette, det er slitsomt, stygg håndskrift og stygge rettelser gir stygge produkter, det er "kjedelig", man må skrive teksten flere ganger. Én av elevene i undersøkelsen skriver i en av tekstene i materialet om egen demotivasjon knyttet til stygg håndskrift (8.2).

Kellogg (1994) er enig i at økt motivasjon gjerne korrelerer med både tekstlengde og tekstkvalitet, blant annet fordi skriveren bruker mer tid på oppgaven; i undersøkelsen i denne avhandlingen er tidsrammen for skrivingen imidlertid temmelig begrenset. Kellogg peker dessuten på at preferanser med hensyn til skriveverktøy er idiosynkratisk, og at man dermed ikke uten videre kan gå ut fra at pc-skriving er motiverende.

3.3.4 Digital skriving

Mange elever opererer på flere skrivearenaer enn før, og mange skriver digitalt både i skoletida og på fritida. I mange av disse sammenhengene er både interaksjonen og produksjonsrammene svært forskjellige fra mange av tekstene som elevene produserer innenfor norskfaget i skolen, og mange elever tar med seg skriveerfaring og kanskje sjangertrekk fra ulike typer digital kommunikasjon (*computer-mediated communication* – CMC) tilbake til de mer monologiske, desituerte skoletekstene.

Kanskje er det slik at en del elever i større grad assosierer de digitale arenaene med tastatur-skriving enn med håndskriving, og at smitteeffekten fra CMC derfor er størst på de tastede tekstene. I og med at mange av de digitale arenaene er preget av mer muntlig preget språkbruk (Baron, 1998; Crystal, 2001; Hård af Segerstad, 2002), er det naturlig å anta at dette kan forsterke de muntlige trekkene i tastetekstene. Trolig vil dette i så fall i størst grad gjelde elever med svakere sjangerforståelse.

Det er forsket mye på språk i digitale diskurser av ulike typer, og ulike sider ved hypertekst og multimodale tekster er også belyst. Imidlertid har jeg funnet lite forskning om språklige variabler i lineære tekster som er fremstilt med digitale verktøy. Crystal (2001) bruker tale/skrift-dikotomien som utgangspunkt når han lanserer begrepet *netspeak* som en "tredje varietet". Han lister opp ulike typer kjennetegn for de to modi og viser hvordan språket på Internett er uensartet og varierer med kommunikasjonskanal, som *chat*, epost, *web*, men låner kjennetegn fra både muntlig og skriftlig språk. Imidlertid blir begrepet *netspeak* dårlig underbygd med hensyn til lingvistiske variabler; det er i hovedsak fundert på variabler knyttet til ytringssituasjonen, som samtaletrekk, språkfunksjoner og mangel på paralingvistiske signaler. Som nevnt i 3.1 ovenfor anser jeg synet på variasjon mellom muntlige og skriftlige diskurser som en en-dimensjonal variasjon for foreldet og grundig tilbakevist.

4 Språkvitenskapelige grunnbegreper

Dette kapitlet definerer og diskuterer kort noen syntaktiske og leksikalske begreper som ligger til grunn for analysene av leksikosyntaktiske variabler i kapittel 9 – 12.

4.1 Syntaktiske begreper

Dette delkapitlet avklarer noen grunnleggende syntaktiske begreper som både korpusbyggingen og språktrekkanalysene bygger på. Tilnærmingen er ikke teoretisk; den står ikke på et spesielt teoretisk fundament, og den er heller ikke et forsøk på å danne et teoretisk fundament. Tilnærmingen er derimot ment å danne en praktisk ramme som i samspill med korpusteknologien og den automatiske morfosyntaktiske annoteringen gjør forskeren i stand til å gjenfinne så mye ulik informasjon fra korpuset som mulig, uavhengig av det teoretiske rammeverket man arbeider innenfor.

Fremstillingen bygger først og fremst på Næs (1965), i mindre grad på Faarlund, Lie og Vannebo (1997), Western (1921) og F.-E. Vinje (1977). En del av terminologien er, så langt jeg kjenner til, original og ikke hentet fra eksterne kilder.

4.1.1 Klausus og subklausus

Det grunnleggende segmentet er *klaususen*, som er en neksus (Næs, 1965, s. 234) med nøyaktig ett subjekt og minst ett finitt verbal i tillegg til eventuelle andre syntaktiske ledd (23). Eventuell innledende (koordinerende) konjunksjon regnes som del av klaususen (24), og det samme gjelder eventuell innledende subjunksjon (25). Subjektet kan bestå av flere koordinerte deler (26), og det kan være flere koordinerte finitte verbaler med neksusforbindelse til samme (eventuelt koordinerte) subjekt (27). Denne definisjonen følger Hunts definisjon av *clause* temmelig nøyaktig (Hunt, 1965, s. 15). Subjektskravet kan tilfredsstilles med relativsubjunksjon, ofte \som\, i subjektsposisjon i klaususen (24). Et unntak fra subjektskravet gjelder klaususer med finitt verbal i imperativ (28).

(23) {De mister rett og slett kontrollen.} [A2-259]
(24) {og som ikke er rettferdig mot noen.} [A1-218]
(25) {at gutter ikke leser bøker,} [A1-202]
(26) {at gutter og jenter leser forskjellige bøker,} [A1-235]
(27) {guttene sitter og "gamer" hele dagen.} [A1-239]
(28) {Følg mitt råd,} [A1-293]

Klaususer kan være av to formelt ulike typer, nemlig *hovedklaususer*, som normalt fungerer som selvstendige syntagmer i en ytring (23), og *subklaususer*, som normalt er underordnet (subordinert) en annen klausus (24) og (25), altså enten hypotaktisk eller ved innføying med Hallidays begreper (3.1). En subklausus kan innledes med en subjunksjon (24) og (25), men ikke nødvendigvis (27). I visse tilfeller kan hovedklaususer også opptre som underordnede klaususer (37), altså som syntaktiske ledd (Diderichsen, 1974[1946], s. 203). Subklaususer kan også fungere som selvstendige pragmatiske enheter i egne ytringer (47). I en

subordineringsrelasjon kan man kalle den overordnede klausus for *overklausus* og den underordnede klausus for *underklausus* i relasjonen.

Der en underklausus fungerer som et obligatorisk ledd i en overklausus, oppstår en terminologisk og begrepsmessig konflikt. Om man ønsker å fokusere på den delen av overklaususen som ikke er inkludert i en underklausus, kan man kalle den en *klaususrest*, jf. Vinjes term "setningsrest" (1977, s. 92), men konseptuelt omfatter overklaususen også ordene i underklaususen. Ordene i underklaususen er dermed del av både underklaususen og overklaususen, i det siste tilfellet som del av ett av leddene i overklaususen (31). Av praktiske grunner knyttet til analysene vil jeg imidlertid normalt regne et ord som medlem av bare én klausus, nærmere bestemt av den subklausus som er nederst i hierarkiet av klaususer som ordet er medlem av. Dette har innvirkning på kvantifiseringen av overklaususen; den blir regnet som kortere og mindre kompleks enn den ellers ville ha vært regnet som, ettersom en obligatorisk del av den ikke er medregnet.

Jeg har til slutt to kommentarer til terminologien jeg har valgt. Så vidt jeg kjenner til, er jeg den første som har brukt termene *klausus* og *subklausus* på norsk; jeg har brukt termen i både undervisning og presentasjoner i flere år, noe som også har inspirert andre (f.eks. Mikkelsen, 2016). Motivasjonen min for denne uvanlige terminologien er polysemien i ordet *setning* i norsk og et behov for et segmentbegrep som omfatter bare én neksus. Når man dessuten som i denne avhandlingen opererer med et studieobjekt som har nær tilknytning til skriveopplæring, blir termen *setning* enda mer problematisk, i og med at den i skolen gjerne brukes om det segmentet som står mellom store skilletegn. Jeg mener derfor at norsk grammatisk terminologi har en lakune og trenger et ord for klausus, tilsvarende *sats* i svensk (Teleman, Hellberg, & Andersson, 1999).

Jeg har valgt termen *klausus* på grunnlag av den engelske betegnelsen *clause*, men begrepene er ikke nødvendigvis de samme. Hunt (1965) og Biber (1988) synes å bruke termen med en betydning som ligger nær mitt klaususbegrep, men Halliday (1987, s. 64) bruker *clause* om både infinitte syntagmer og subjektsløse syntagmer: Dette innebærer at Hallidays analyser ikke er direkte sammenlignbare med mine. Chafe og Danielewicz (1987, s. 95) baserer derimot sine analyser på det de kaller en *intonation unit* i stedet for et klausus- eller *clause*-begrep, noe som virker fornuftig ut fra at deres studieobjekt omfatter muntlige ytringer. Deres analyser blir dermed heller ikke direkte sammenlignbare med mine.

4.1.2 T-enhet

Klaususer opptrer gjerne ikke alene, men i kombinasjoner med hverandre, og vi trenger en mer overordnet segmenttype som bedre kan fange kompleksiteten i forskjellige typer ytringer. Hunt var interessert i syntaktiske trekk som uttrykk for elevenes utvikling i modenhet og pekte på at *setningen* (*sentence*), definert som det som står mellom to store skilletegn, ikke er særlig egnet til dette, fordi elever har ganske ulike vaner når det gjelder tegnsetting (1965, s. 6-12). Han nevner som et eksempel en fjerdeklassing (*fourth grader*) som skrev hele sin 77 ord lange tekst uten skilletegn og dermed oppnådde en gjennomsnittlig setningslengde på 77 ord. Elevene fra VG1 i min undersøkelse har ikke et like arbitrært eller

anarkistisk forhold til tegnsetting, men det er liten tvil om at tegnsettingsvanene deres til dels er såpass tilfeldige og i hvert fall såpass varierende at tegnsetting er lite egnet som grunnlag for analyseenheter (32), med mindre det er tegnsettingen i seg selv som er studieobjektet.

Hunt foreslår (s. 20-) på bakgrunn av dette den analyseenheten han betegner en *t-unit*, som en forkortelse for "*minimal terminable unit*", definert som "en hovedklausus pluss de subordinerte klausale eller ikke-klausale strukturer som er tilføyd eller innføyd i den" (Hunt, 1970, s. 4). På norsk bruker vi termen t-enhet (f.eks. Jensen & Steien, kommer). Begrunnelsen for betegnelsen er basert på at t-enhetene er de minste enhetene en tekst kan deles opp i med store skilletegn ved å følge tradisjonelle tegnsettingsregler. I et annet perspektiv er t-enheten også en maksimal enhet, i den forstand at man når man deler en tekst inn i t-enheter, søker å få alt språklige materiale i teksten inn i en (og akkurat én) t-enhet.

T-enheter kan ha hovedklaususer som er fortellende, spørrende eller med finitt verbal i imperativ (29), (34), (35), (36).

Eksemplene (29) – (36) nedenfor illustrerer t-enheter av ulike typer. (32) viser et eksempel på en t-enhet som overskrider store skilletegn, mens (33) viser grafiske setninger som overskrider t-enheter. (30) viser et eksempel på ikke-klausalt materiale (\Nei\) som tilordnes t-enheten.

(29) {$_{T}$ Foreldre gir barna sine alkohol} [A2-259]
(30) {$_{T}$ Nei, jeg tror ikke det.} [A2-259]
(31) {$_{T}$ Min mening er {S at butikkene bare er ute etter penger.}} [A2-259]
(32) {$_{T}$ Jeg bruker mye tid foran datamaskinen selv, og snakker ofte med venner og bekjente over internett etter skoletid. Både gutter og jenter.} [A1-297]
(33) {$_{T}$ Men man mister kondis, muskler,} {$_{T}$ og leveren blir ødelagt over lang tid med drikking.} [A2-239]
(34) {$_{Timp}$ Gjør det {S som passer deg,}} [A1-264]
(35) {$_{Tspm}$ Hvorfor er visse jobber og stillinger kjønnsdelt?} [A1-203]
(36) {$_{Tspm}$ Er dette riktig da?} [A1-203]

Hunt illustrerer med eksempler hvordan definisjonen brukes i praksis, men tilbyr ikke noen retningslinjer som eksplisitt avklarer tvilstilfeller. Sotillo (2000, s. 109) gir eksplisitte retningslinjer, men ser ut til å støtte seg mer på tegnsetting enn Hunt. Også Foster (2000, s. 360) nevner arbeider som delvis støtter seg på tegnsetting i segmentering av tekst i t-enheter. Med støtte delvis i eksemplene i Hunt og delvis i enkelte av retningslinjene fra Sotillo har jeg konstruert noen mer entydige retningslinjer:

- Direkte tale og andre sitater blir segmentert som t-enheter på linje med annen tekst, såfremt sitatet ikke inngår som ledd i en overklausus (41).
- Dersom en t-enhet fungerer som ledd eller ledd-del i en overklausus, typisk direkte objekt, regnes den som del av t-enheten som overklaususen er en del av. Slike subordinerte t-enheter er ofte sitater, men det kan også dreie seg om andre typer formuleringer, som i (37) og (39).

- Dersom et sitat eller lignende inneholder mer enn én hovedklausus, regnes den første som del av overklaususen, men alle påfølgende hovedklaususer som kjerner i selvstendige t-enheter (38).
- Tekst i parenteser analyseres som annen tekst, med ett unntak: En t-enhet kan ikke inneholde en annen t-enhet, så parenteser som står inni en t-enhet, må regnes som en del av den omsluttende t-enheten.
- En koordinert subklausus med elidert subjunksjon regnes som del av foregående t-enhet, selv om den koordinerte subklaususen har form som en hovedklausus (42).

En del konstruksjoner er formelt tvetydige, og man må støtte seg på en semantisk tolkning for å avgjøre om en koordinert klausus er en underordnet subklausus eller en selvstendig hovedklausus.

(37) {T Mitt spørsmål er da: {S Er det virkelig så enkelt?}} [A2-261]
(38) {T og i denne artikkelen skriver redaktøren {S «Hva gjør foreldrene {S når 14-15-åringene kommer og ber om en øl til kveldens fest?}}} {T Svaret bør være enkelt.»} [A2-210]
(39) {T {S kommer da aldri noe bra ut av det} mener nå jeg.} [A2-285]
(40) {T Påstanden {S «Gutter leser ikke bøker, men driver med data.} {S Jenter driver ikke med data, men leser bøker»} er i mine øyne basert på forskning {S som viser {S at jenter gjør det bedre på skolen enn gutter.}}} [A1-269]
(41) {T Jeg henviser meg til dette sitatet:} {T «Gutter leser ikke bøker, men driver med data.} {T Jenter driver ikke med data, men leser bøker.»} [A1-205]
(42) {T Det virker {S som du mener {S at gutter ikke er ment til å lese} {S og jenter ikke er ment til å drive med teknologi.}}} [A1-292]

I noen få tilfeller inneholder tekstene konstruksjoner der flere t-enheter sammen fungerer som et syntaktisk klaususledd (43). I disse tilfellene har jeg segmentert hver underordnede t-enhet som en subklausus; det finnes også andre eksempler på at jeg har måttet velge en praktisk tilnærming til segmentering av et materiale som ikke i alle henseende følger vanlige konvensjoner for skriftlig språkbruk, for eksempel (44) og (45).

(43) {T {S Hvis en 15-åring sender med en femhundrelapp og sier: {S «Her, ta denne du,} {S og så kjøper du drikke for alle penga.»}} så er det selvfølgelig {S at noen gjør det.}} [A2-205]
(44) {T Jeg mener {S vi må huske på {S at det er mye nyttig å lære på data,} {S og at {S en bok kanskje gir deg nyttige norsk- og skrivekunnskaper,} kan også en data gjøre det.}}} [A1-212]
(45) {T Dette er jo bare tull} {T å komme med et slikt argument uten noen fornuftige kilder å henvise til blir bare dumt.} [A1-215]

I noen ganske få tilfeller, totalt 50 segmenter i 120 tekster, har jeg justert definisjonen av t-enheter til å omfatte segmenter som ikke har en neksus-basert hovedklausus i kjernen. Se diskusjonen av dette i 4.1.3 nedenfor.

4.1.3 Fragment

Som nevnt over kan t-enheten sees som en maksimal enhet i den forstand at en segmentering av en tekst i t-enheter skal omfatte alt språklig materialet i teksten. Det innebærer at eventuelle analyserester skal forsøkes å tolkes som en ikke-klausal struktur som er tilføyd eller innføyd i hovedklaususen, og denne tolkningen skal være uavhengig av tegnsetting.

I praksis er det ikke alltid mulig å tolke alt materiale på denne måten med noen slags gyldighet. Slik forskere på muntlige ytringer har funnet (f.eks. Foster, Tonkyn og Wigglesworth, 2000, s. 360), finnes det fragmenter som det ikke formelt eller innholdsmessig er mulig å tilordne en overklausus, uavhengig av tegnsettingen i teksten. I totalt 80 tilfeller i korpuset på 120 tekster har jeg funnet det nødvendig å utelate klaususfragmenter fra en t-enhet. Disse tilfellene skiller seg i to ulike typer som jeg har behandlet på prinsipielt ulik måte.

For det første finnes det fragmenter som inneholder (minst) et finitt verbal, men som enten ikke omfatter noen hovedklausus (47), eller der det som kunne ha vært en hovedklausus ikke har subjekt (46) eller har andre typer mangler som gjør at det ikke tilfredsstiller kravene til en hovedklausus (48). Disse segmentene har kompleksitet som hovedsakelig svarer til kompleksiteten i klaususer og t-enheter, og jeg har valgt å kalle dem *finitte fragmenter* og regne dem med blant t-enhetene. I korpuset er disse tagget som `<t-unit type="frag">`. Det finnes totalt 50 slike finitte fragmenter i korpuset.

(46) {Tfrag Sier heller ikke det {S at gutter ikke chatter,}} {Tfrag men tror nok {S gutter bruker mere av tiden på data til å spille enn å chatte.}} [A1-303]
(47) {Tfrag {S fordi data er den nye verden.}} [A1-237]
(48) {Tfrag Nei, så hvorfor slenge ut en sånn påstand, {S som ikke kan stemme i det hele tatt?}} [A1-233]

For det andre finnes det klaususfragmenter som ikke inneholder noe finitt verbal, og dem kaller jeg ekte fragmenter. De er tagget i korpuset som `<frag>` og regnes altså ikke som t-enheter. Se (49) – (52) nedenfor. Det finnes totalt 30 slike ekte fragmenter i korpuset; 7 av disse står i den samme teksten ([A1-312]).

(49) {F Men hva med andre bøker?} [A1-205]
(50) {F Tilbake til påstanden.} [A1-260]
(51) {F og gudskjelov for det.} [A1-203]
(52) {F Nei?} [A1-312]

Noen ekte fragmenter fungerer som diskursmarkører i teksten og kan for eksempel signalisere overgang til et nytt tema, som for eksempel (50).

4.1.4 Subklaususkategorier

Når man skal kategorisere subklaususer, kan man bruke indre, formelle kriterier eller ytre, funksjonelle kriterier. Siden de formelle kriteriene uansett i stor grad er søkbare i korpuset,

har jeg valgt å kategorisere subklaususene etter funksjonelle kriterier. Jeg følger Næs' (1965) ganske tradisjonelle tredeling i nominale, adverbiale og relative klaususer. Ettersom analysene i kapittel 11 ikke tematiserer andre kategorier enn den adverbiale, fokuserer denne diskusjonen mest på denne typen subklaususer.

Nominale subklaususer fungerer som nominale ledd i overklaususen. De er gjerne innledet med subjunksjonen \at\, subjunksjonen \om\ eller spørreord som fungerer som subjunksjon, men alle disse innlederordene kan også fungere som subjunksjoner i andre kategorier av subklaususer. \om\ kan også innlede adverbiale vilkårsklaususer, mens både \om\, \at\ og spørreord kan innlede relativklaususer. I visse omgivelser er innledende \at\ ikke obligatorisk, og en del nominale subklaususer er derfor subjunksjonsløse.

I tillegg til nominale subklaususer kan hovedklaususer fungere som nominale ledd (4.1.1). I korpuset og i analysene er disse kategorisert som subklaususer, men som en egen kategori adskilt fra de nominale subklaususene. Dette kan muligens sees på som en brist i systematikken i tilnærmingen, men det har ingen praktiske konsekvenser i analysene. Underklaususer som har hovedklaususform, blir altså regnet som subklaususer i analysene.

Relative subklaususer fungerer ikke som egne ledd i overklaususen, men som en del av et ledd. De innledes oftest av \som\, \at\ eller \om\; \som\ er ikke obligatorisk i alle omgivelser, og en del relative subklaususer er derfor subjunksjonsløse. En subkategori er relative tidsklaususer, som har innlederord som \som\, \da\, \når\, \etter at\; også for disse er subjunksjonen i visse omgivelser ikke obligatorisk. Dessuten finnes relative stedsklaususer, som har innlederord som \som\, \der\, \hvor\, \der som\, \der hvor\ (54). En annen subkategori er relative sammenligningsklaususer, typisk innledet av \enn\ eller \som\ i konstruksjoner av typen i \større enn\ eller \så stor som\, der adjektivet fungerer som korrelat i overklaususen. Både relative tidsklaususer, relative stedsklaususer og relative sammenligningsklaususer kan være vanskelige å skille fra adverbiale subklaususer med tilsvarende funksjoner. Til slutt må det nevnes at det finnes relative subklaususer der korrelatet i overklaususen er en neksus eller et allment \den\, \det\, \de\, \hvem\ eller \hva\.

Adverbiale subklaususer fungerer som adverbiale ledd i overklaususen. De adverbiale leddene kan være av ulike kategorier, med funksjoner knyttet til tid (53), sted (54), årsak (55), betingelse (56), hensikt (57), følge (58), innrømmelse (59) og sammenligning (60). Dessuten finnes adverbiale subklaususer med en adversativ funksjon (61) og enkelte tilfeller som ikke har noe klart adverbialt innhold men kan ha en tekstordnende rolle (62) og (63). Alle de tre siste eksemplene har form som temporale subklaususer, men deres semantiske betydning eller pragmatiske rolle i teksten er neppe temporal.

(53) {T Og ja, for mange prektige og staselige foreldre {Srel som "aldri" gjorde noe galt {Sadv-tid da de var unge,}} er det sikkert det.} [A2-312]

(54) {T De vil jo være {Sadv-sted der det skjer,} og ikke ute i skogen {Srel-sted der det ikke skjer.}} [A2-265]

(55) {T Jeg tror {Snom at guttene bruker litt mer tid enn jentene,} {Sadv-årsak fordi gutten spiller,}} [A1-213]

(56) {T og {Sadv-betingelse dersom du vil passe inn,} må du følge de normene {Srel folk rundt deg setter opp.}} [A1-227]

(57) {T «Vi gir ungen drikke {Sadv-hensikt så vi vet {Snom hva han / hun drikker.»}}} [A2-211]

(58) {T Forhåpentligvis har titusener blitt spurt, {Sadv-følge så svarene har blitt forskjellige,}} [A2-272]

(59) {T {Sadv-innrømmelse Selv om de er 12 eller 16 år,} så hører de ikke på foreldrene.} [A2-238]

(60) {T og det ser ikke ut {Sadv-sammenligning som om at det kommer til å roe seg ned med det første.}} [A2-252]

(61) {T for ene fordelen med internett er {Snom at det oppdaterer seg hele tiden,} – {Sadv mens blader og bøker tar en liten stund {Sadv før det blir noe nytt.}}} [A1-208]

(62) {T Så jeg mener da {Snom at det er ganske likt {Sadv når det gjelder gutter og jenter og data og bøker.}}} [A1-202]

(63) {T {Sadv Når jeg tenker meg om,} så er det kanskje litt dumt,} [A1-202]

Skillet mellom for eksempel stedsklaususer av relativ og adverbial type er altså prinsipielt om det finnes et korrelat i overklaususen. I noen tilfeller kan imidlertid et slikt korrelat være utydelig, og markører som \der som\ eller \der hvor\ kan analyseres som korrelat etterfulgt av subjunksjon eller som én subjunksjon bestående av to ord. Det finnes bare ett slikt eksempel i korpuset (64), og jeg har analysert det som korrelat pluss subjunksjon. Ord som \der\ og \da\ kan fungere både som korrelat og som subjunksjon, som illustrert av (54).

(64) {T Men jeg har òg vært der {Srel hvor ingen drakk,} {Srel alle hadde det gøy,} {Srel ingen satt og var utenfor,} {Srel og ingen hadde følelsen av {Snom at de ikke var kule nok,} eller følelsen av å være usosial.}} [A2-233]

\Fordi\ og \mens\, som tradisjonelt kategoriseres som subjunksjoner (Faarlund, et al., 1997, s. 88, 1067), er semantisk nært beslektet med konjunksjonene \for\ og \men\. \For\ og \fordi\ uttrykker kausalitet (i samme retning), mens \men\ og ofte \mens\ uttrykker en motsetning (61). \Mens\ er polysemt og har dessuten en variant med temporal betydning. (Faarlund, et al., 1997, s. 1067) I tillegg til den semantiske nærheten er det en åpenbar fonologisk nærhet mellom \for\ og \fordi\ på den ene siden, og \men\ og \mens\ på den andre. Begge disse likhetstrekkene spiller trolig en rolle når vi ser at subjunksjonene ofte tar de syntaktiske rollene til sine beslektede konjunksjoner. Når \fordi\ og \mens\ innleder klaususer som ikke bare har helsetningsplassering av setningsadverbial, men også adverbialer eller andre ledd i forfelt, eller når de til og med innleder lengre segmenter av flere helsetninger, er det naturlig å vurdere om det har skjedd en rekategorisering av ordene, og at de nå fungerer som konjunksjoner. Dette kan i så fall ha konsekvenser for segmenteringen i t-enheter. Imidlertid har jo ikke ordene mistet sin subordinerende funksjon, og ettersom hovedklaususer og subklaususer ikke nødvendigvis skiller seg fra hverandre formelt, kan det være prinsipielt umulig å avgjøre om \fordi\ eller \mens\ fungerer koordinerende eller subordinerende i et gitt tilfelle.

En tilsvarende potensiell konflikt gjelder \for\, som tradisjonelt kategoriseres som konjunksjon, men som i praksis også kan fungere som subjunksjon. Oslo-Bergen-taggerprogrammet (se 6.4.2), som er brukt til den automatiske lemmatiseringen og morfologiske annoteringen av tekstmaterialet, følger tradisjonen for disse tre ordene og kategoriserer alltid \fordi\ og \mens\ som subjunksjoner og \for\ som konjunksjon. For å unngå forstyrrende konflikter mellom den automatiske og den manuelle taggingen har jeg valgt også å følge tradisjonen her, og kategoriserer \fordi\ og \mens\ alltid som subjunksjoner og deres etterfølgende klaususer som subklaususer, og \for\ alltid som konjunksjon.

4.2 Leksikalske begreper

Mens enkelte av de syntaktiske begrepene jeg benytter, er litt utradisjonelle i norsk språkvitenskap, er de leksikalske begrepene i avhandlingen stort sett ukontroversielle og uten behov for noen diskusjon. Et unntak er skillet mellom leksikalske og grammatiske ord, som diskuteres i 9.2.1. Begrepene preposisjon og attributivt adjektiv avklares også i kapitlene som analyserer variablene knyttet til disse fenomenene, henholdsvis 11.2.1 og 11.3.3.

Fordi korpusløsningen som er benyttet i prosjektet, har støttet seg på ulike versjoner av lemmatiseringsprogrammet i løpet av prosessen (se 6.4.2 og 8.4.1), brukes to ulike definisjoner av ord. I den første versjonen av programmet (CG1) er definisjonen av ord basert på det grafiske ordet, mens den andre versjonen av programmet (CG3) tillater enkelte flerordsleksemer.

Lemmatiseringsprogrammet tilegner en lemmaform og en ordklasse til hvert ord, men analysen resulterer ikke alltid egentlig i en entydig tilknytning til leksem. Derfor forholder analysene seg kun til to nivåer av orddefinisjon, nemlig ordformen, slik den står i korpusteksten, og lemmaformen som lemmatiseringsprogrammet gir.

Når det gjelder morfologiske opplysninger om hvert ord, benytter analysene de kodene som den automatiske taggeren produserer. I denne avhandlingen er det først og fremst ordklasseinformasjon som er relevant, i tillegg til finitthet ved verbalene. Jeg drøfter undervegs i avhandlingen de tilfellene jeg har funnet det nødvendig å justere analysene fra taggerprogrammet.

5 Hypotese og forskningsspørsmål

Hypotesene i denne undersøkelsen er knyttet til hovedsakelig to faktorer: skrivehastighet og redigeringsmuligheter. Begge faktorene forutsetter at pc-ferdighetene til elevene er tilstrekkelige til at de (1) skriver raskere på tastatur enn for hånd og (2) redigerer lettere i tekstbehandlingsprogram enn på papir. Hypotesen gjelder altså bare for elever med slike ferdigheter. Effektene av at tekstene er produsert i tekstbehandlingsprogram, vil trolig være annerledes for elever med dårligere pc-ferdigheter, og tekster fra slike elever er forsøkt holdt utenfor undersøkelsen. (Se 6.1.2 om utvalg av elever.) Det er rimelig å anta at det også er en forutsetning for hypotesene at de (3) har en positiv holdning til å arbeide med pc-verktøy.

5.1 Overordnet hypotese

Den overordnede hypotesen er på bakgrunn av diskusjonen over treleddet:

- Større skrivehastighet gir flere spontane ("muntlige") trekk i tekstene.
- Bedre redigeringsmuligheter gir flere planlagte ("skriftlige") trekk i tekstene.
- I skrivesituasjoner der hastighetsfaktoren har størst gjennomslag, vil tastede tekster ha flere spontane trekk enn håndskrevne, mens i skrivesituasjoner der redigeringsfaktoren har størst gjennomslag, vil tastede tekster ha flere planlagte trekk enn håndskrevne.

Jeg antar videre at flere parametre har innvirkning på hvilken faktor som har størst gjennomslagskraft i hver enkelt tekst.

Tidsrammen spiller trolig en rolle. Korte tidsrammer kan favorisere hastighetsfaktoren, men kan også favorisere redigeringsfaktoren ved at det blir forholdsvis mer tid tilgjengelig til redigering når man bruker det raskeste verktøyet.

Elevens ferdigheter virker antagelig inn. Elever som er flinke til å nyttiggjøre seg omskrivingsfasene i skriveprosessen, vil produsere tekster med flere planlagte trekk på tastatur.

Oppgaven er nok ikke uvesentlig. Oppgaver som skaper motivasjon for å arbeide med en tekst over lengre tid, favoriserer redigeringsfaktoren. Oppgaver som skaper personlig engasjement, kan favorisere hastighetsfaktoren, men kan også favorisere redigeringsfaktoren gjennom økt motivasjon for redigering.

Teksttype eller sjanger kan være en faktor. Muntlig pregede sjangre eller registre, som fortellinger, favoriserer hastighetsfaktoren, både gjennom at produksjonen kan ha en tendens til å gå raskere, og gjennom at idealet for teksttypen vil ha et mer spontant preg.

5.2 Diskusjon

Den endringen i prosesseringsbetingelser jeg søker å undersøke, kan påvirke skriveren på andre måter enn de to faktorene jeg baserer hypotesen på: hastighet og redigering.

Motivasjonen kan endres. Elever kan anspores til å velge andre sjangre eller stilnivå. Det presset som skoleelever opplever for å skrive "langt nok", enten dette er selvpålagt eller lærerpålagt, fungerer trolig ulikt på papir og skjerm. Det er neppe mange elever som har nøyaktig intuisjon om hvor mange ord deres håndskrevne tekster består av, eller hvor mange sider i Times New Roman 12 de tilsvarer.

De ulike faktorene jeg nevner ovenfor, kan også interagere med hverandre og forsterke eller påvirke hverandre. For eksempel vil elever med sterk sjangerbevissthet og gode ferdigheter i omskriving kunne produsere tekster med flere spontane trekk dersom dette er idealet for den sjangeren de skriver i. Engasjement kan også i sin tur skape motivasjon for å arbeide mer med teksten. Motivasjon korrelerer gjerne med både tekstlengde og tekstkvalitet (Kellogg, 1994), men hvilken innvirkning dette vil ha på det syntaktiske nivået, er mer usikkert. Det er også mulig at skriveverktøyet kan påvirke skriverens valg av sjanger og register og på den måten forrykke tekstens syntaktiske egenskaper.

Dertil kommer at det ikke vil være noen en-til-en-relasjon mellom aspekter ved skrive-situasjonen og lingvistiske trekk i teksten, noe som også reflekteres av den mangelen på funn av absolutte forskjeller mellom skrift og tale som preger tidligere forskning (kapittel 3). En kan derfor ikke vente seg entydige data på overflaten, men en flerdimensjonal variasjon som kan avhenge av at trekkene analyseres på riktig detaljeringsnivå; for eksempel kan en analyse av antall leddsetninger totalt gi et annet resultat enn antall adverbiale leddsetninger, eller antall leksikalske ord per løpeord kan gi et annet resultat enn antall leksikalske ord per klausus. Bibers advarsel (1986, s. 384) mot å legge for stor vekt på enkelte trekk, individuelle tekster og individuelle teksttyper tilsier dessuten bruk av et materiale som er bredt nok med hensyn til

- antall språklige variabler
- antall tekster
- antall teksttyper
- antall skrivere

I dette forsøket er antall tekster og antall skrivere ganske lavt, og variasjonen i teksttyper er bevisst holdt så liten som mulig nettopp med tanke på å nøytralisere denne faktoren. Det vil si at undersøkelsens resultater prinsipielt ikke kan generaliseres til andre teksttyper.

5.3 Forskningsspørsmål

Hypotesen i 5.1 forsøker å predikere om spontane versus planlagte trekk, men nevner ikke hva slags trekk som kan sies å være henholdsvis spontane og planlagte. Gjennomgangen av

teori og tidligere empiri om kompleksitet, register og skriving i kapittel 2 og 3 gir visse indikasjoner om hva slags språklige variabler som er aktuelle, men få konkrete variabler.

Variabler som peker seg ut som interessante å analysere, er særlig slike som Biber (1986, 1988) påviser som *fremtredende* (*saliente*) i sin tekstdimensjon 1, interaktivitet kontra redigeringsgrad, i hvert fall de av dem som er overførbare til norsk. Et mer pragmatisk kriterium er at variablene må kunne la seg trekke ut av korpuset uten for mye manuell analyse. Hallidays todimensjonale perspektiv på kompleksitet som sammensatt av innfløkthet og kompakthet oppfatter jeg også som opplysende, og Bibers resultater og Hallidays spekulasjoner både utfyller og bekrefter hverandre i stor grad. Disse to tilnærmingene til register og variasjon danner det viktigste grunnlaget for utvalg av konkrete variabler og diskusjon av resultater, men også Vagles (1990) undersøkelse av språkvariabler i radiospråk er relevant, og variabler som andre forskere knytter til register (Chafe, 1982; og resultater som er referert av Chafe & Tannen, 1987; Halliday, 1979) eller til tekstkvalitet (Russell, 1999; Russell & Haney, 1997; Vagle, 2005b) er interessante.

5.3.1 Leksikosyntaktiske variabler

Forskningsspørsmålene er knyttet til to typer av språklige variabler, syntaktiske variabler forankret i t-enheten og leksikalske variabler knyttet til variasjon og informasjonell tetthet. Dessuten er forskningsspørsmålene knyttet til 4 elevfaktorer, i tillegg til skriveverktøyet.

Variablene blir utviklet og diskutert i de tre analysekapitlene 9, 10 og 11, og disse kapitlene drøfter også en del potensielle variabler som blir forkastet som mindre aktuelle av ulike grunner. Jeg oppsummerer her konklusjonene av disse drøftingene ved å liste opp de 13 leksikosyntaktiske variablene som jeg vurderer som mest verdifulle. 10 av disse brukes i en helhetlig prinsipalkomponentanalyse av det overordnede forskningsspørsmålet i kapittel 5.1:

- gjennomsnittlig ordlengde
- gjennomsnittlig ordlengde i leksikalske ord
- leksikalsk tetthet
- globalt type/eksemplar-forhold (TTR)
- lokalt type/eksemplar-forhold (TTR)

- gjennomsnittlig t-enhetslengde
- andel t-enheter med korte forfelt
- gjennomsnittlig klaususlengde
- subklaususfrekvens
- andel adverbiale subklaususer
- frekvens av korte subklaususer
- preposisjonsfrekvens
- frekvens av attributive adjektiver

De konkrete forskningsspørsmålene blir dermed utviklet undervegs i analysekapitlene, og andre kandidater til leksikosyntaktiske variabler er utelukket på bakgrunn av ulike typer argumenter: praktiske, tekniske, matematiske eller validitetsmessige.

5.3.2 Faktorer / prediktorer

Av den store mengden av teoretisk mulige påvirkningsfaktorer har jeg i analysene gjort et utvalg av 4 elevfaktorer i tillegg til skriveverktøyet:

- kjønn (gutt, jente)
- skriveferdigheter (middels, sterk)
- total tekstlengde (kort, lang)
- forskjell i tekstlengde (liten, stor)

Det er 30 gutter og 30 jenter i utvalget.

Middels skriveferdighet er definert som karakteren 3 eller 4 som standpunktkarakter fra ungdomsskolen, mens sterk skriveferdighet representerer karakteren 5 eller 6.

Total tekstlengde er en dikotom variabel basert på medianen av summen av antall ord i hver tekst. Elever som har total tekstlengde lavere enn medianen, får faktorverdien "kort", mens resten får verdien "lang".

Forskjell i tekstlengde er en dikotom variabel basert på medianen av kvotienten av antall ord i tasteteksten og antall ord i håndteksten, altså tastetekstens lengde delt på håndtekstens lengde. Elever som har verdier lavere enn medianen, får faktorverdien "liten" forskjell, mens resten får verdien "stor" forskjell.

Dette innebærer at alle de fire faktorene er dikotome, og at for hver faktor er det 30 elever i hver kategori. Det er liten eller ingen interaksjon mellom de 3 faktorene kjønn, skriveferdighet og forskjell i tekstlengde, mens total tekstlengde interagerer noe med kjønn og en del med skriveferdighet og forskjell i tekstlengde. Se 7.3.2.2 for en diskusjon av dikotomisering av tekstlengdevariablene.

Selv om det ikke er interaksjon mellom kjønn og skriveferdighet, er det noe interaksjon mellom kjønn og norskkarakter. De 2 elevene som har karakteren 3 i norsk, er begge gutter, og 4 av de 5 elevene som har karakteren 6, er gutter. Antallet elever som faller utenfor de normale karakterene 4 og 5, er imidlertid såpass lavt at det neppe påvirker resultatene særlig mye.

Elevenes egenskaper er presentert og drøftet mer inngående i 6.1.

5.3.3 Oppsummering

Basert på diskusjonen ovenfor om hvordan skriveverktøyet kan påvirke språktrekk i spontan eller planlagt retning, og hvordan ulike faktorer innvirker på skriveverktøyets påvirkning på

språktrekkene, har jeg formulert følgende generiske forskningsspørsmål for denne undersøkelsen:

> Blir variabel X påvirket i spontan eller planlagt retning av skriveverktøyet, og arter denne påvirkningen seg ulikt for ulike delsegmenter av elever?

Metode

De neste to kapitlene presenterer og diskuterer metoder for datainnsamling og analyse. Det første kapitlet presenterer datainnsamlingen. Det inkluderer et delkapittel om korpusteknologien og byggingen av korpuset. Det neste kapitlet presenterer og drøfter metodene for de statistiske analysene som er brukt i avhandlingen.

6 Datainnsamling

Dette kapitlet beskriver prosedyrene rundt rekruttering av informanter, innsamling av tekster og innhenting av annen type informasjon om elevene og tekstene.

Den grunnleggende metodologiske ideen i eksperimentet er en paret design som danner grunnlag for å studere tekstpar. To og to tekster er skrevet av samme skribent, hvorav én tekst for hånd og én på pc-tastatur med tekstbehandlingsverktøy på en datamaskin. Alle elevene skrev tekster som var besvarelser på de samme to oppgavene, og gruppen ble delt slik at halvparten av elevene skrev den første oppgaven for hånd og den andre på tastatur, mens den andre halvparten skrev den første oppgaven på tastatur og den andre for hånd, en såkalt ABBA-design. På denne måten er materialet i best mulig grad sikret mot tilfeldige skjevheter i fordelingene av parametre.

6.1 Elever

Mitt første initiativ til å kontakte potensielle informanter var ved å ta kontakt med rektorer og lærere i aktuelle videregående skoler i forskjellige deler av landet. Det viste seg i praksis å være vanskelig å gjøre avtaler med skoler og lærere, inntil jeg fikk kontakt med en interessert fagansvarlig lærer ved en videregående skole i et tettsted på indre Østlandet. Kontakten med akkurat denne skolen oppstod litt tilfeldig ved at jeg diskuterte prosjektet med læreren i en annen sammenheng, hvorpå læreren syntes prosjektet var interessant og foreslo å legge det frem for de andre norsklærerne ved skolen. Etter den fagansvarliges orientering til de andre lærerne hadde vi et fellesmøte mellom alle norsklærerne og meg, og vi ble enige om at de fire norsklærerne skulle rekruttere informanter og samle inn tekster fra sine respektive klasser.

Datainnsamlingen fant sted i vårsemesteret 2009, og alle informantene i denne undersøkelsen var på det tidspunkt elever på studiespesialiserende utdanningsprogram i VG1 ved én videregående skole i et tettsted på indre Østlandet. Avtaler ble gjort med 105 elever, hvorav 101 av elevene fylte ut spørreskjemaet (se 6.1), der de blant annet fylte ut opplysninger som var viktige parametre for utvalg av informanter. 94 elever gjennomførte prosjektet etter forutsetningene.

6.1.1 Spørreskjema

Alle elevene fylte ut et kort spørreskjema før skriveøktene. Dette skjemaet hadde tre hovedfunksjoner. For det første skulle skjemaet gi bakgrunnsinformasjon som grunnlag for utvelgelse av elever med egenskaper som er i tråd med forutsetningene for eksperimentet. For det andre skulle skjemaet gi bakgrunnsinformasjon som grunnlag for balansering av visse faktorer knyttet til elevenes egenskaper (6.1.2), nemlig kjønn og skriveferdighet. For det tredje skulle informasjon om elevenes holdninger til medier og hvordan de bruker forskjellige medier, danne grunnlag for elevparametre som kunne bidra i analyse og tolkning av resultatene. Det siste momentet er i liten grad utnyttet i undersøkelsen, blant annet fordi elevene klumpet seg i enkelte av svaralternativene, noe som gjorde det vanskelig å utnytte

informasjonen i hypotesetestende statistikk. Men i noen grad er informasjonen benyttet i tolkning av resultatene.

I løpet av datainnsamlingsperioden ble det klart for meg at spørreskjemaet hadde fokusert for ensidig på elevenes bruk av pc og holdninger til bruk av pc, og jeg sendte derfor ut et supplerende spørreskjema i etterkant av tekstinnsamlingen. Dette skjemaet fokuserte i større grad på håndskriving og holdninger til håndskriving. Jeg kaller de to skjemaene henholdsvis Skjema A og Skjema B. Begge skjemaene er gjengitt i sin helhet i henholdsvis appendiks B6 og appendiks B7. Begge skjemaene var på papir og krevde avkrysning og noe skriving av stikkord.

Skjema A var på fire sider, og alle deltagende elever fylte ut dette skjemaet. Spørsmålene var nummerert, og de fleste hadde avkrysningsmuligheter for ferdige svaralternativer. Noen av spørsmålene krevde avkrysning for bare ett svaralternativ, mens andre tillot flere. "Vet ikke" var ikke et svaralternativ, men ettersom skjemaet var på papir, var det selvfølgelig mulig å la være å krysse av for enkelte spørsmål. Elevene hadde generelt god skjemadisiplin, og det var veldig lite blanke svar. Skjemaet bestod av tre hoveddeler: "dine holdninger til pc", "din bruk av pc" og "personlige opplysninger".

Det var tre hovedspørsmål om elevens holdninger til pc:

- Egen beherskelse av tekstbehandlingsprogram på pc
  ("bra", "ganske bra", "ganske dårlig", "dårlig")
- Foretrekker pc eller håndskriving til skolerelaterte skriveoppgaver
  ("alltid pc", "oftest pc", "oftest hånd", "alltid hånd")
- Fordeler ved det foretrukne skriveverktøyet. 8 ferdigformulerte alternativer
  ("viktig", "litt viktig", "ikke viktig", "usant")

En svakhet ved det siste spørsmålet er at det ikke etterspurte fordeler (eller ulemper) ved begge de alternative verktøyene, men bare det foretrukne. Det var inget åpent alternativ til dette spørsmålet.

Det var fire hovedspørsmål om elevens bruk av pc.

- Hvilke typer aktiviteter eleven bruker pc til utenom skolen. Svaralternativene var utformet særlig med tanke på å avdekke om eleven bruker pc mye til skriveaktiviteter. Spørsmålet har avkrysningsalternativer, men også to åpne spørsmål der eleven kan skrive hvilke spill han eller hun bruker, og andre typer aktiviteter som ikke er dekket av de spesifiserte alternativene.
  ("mye", "en del", "lite", "aldri")
- Hvor ofte eleven bruker pc utenom skoletid – både til skolearbeid og fritidsaktiviteter. ("Hver dag eller nesten hver dag", "1 – 4 dager i uka", "Sjeldnere enn 1 dag i uka")
- Hvor lenge eleven bruker pc utenom skoletid en typisk dag.
  ("Over 2 timer", "1/2 - 2 timer", "Under 30 minutter")

- Egen vurdering av vanskelighetsgrad for hver av 12 nevnte redigeringsaktiviteter i tekstbehandlingsprogrammet
    - skrive inn tekst
    - rette feiltastinger
    - sette inn et nytt ord i en setning
    - sette inn en ny setning
    - sette inn et nytt avsnitt
    - slette et ord i en setning
    - slette en setning
    - slette et avsnitt
    - flytte et ord i en setning
    - flytte en setning
    - flytte et avsnitt
    - bruke angrefunksjonen

  ("lett" eller "litt vanskelig")

Til slutt var det en kort del som innhentet opplysninger om kjønn, morsmål, hovedmål og standpunktkarakter i norsk hovedmål fra ungdomsskolen. For opplysningen om morsmål var det bare to alternativer, nemlig "norsk" og "annet", men det var eksplisitt opplyst at man kunne krysse av for "ett eller flere alternativ".

Standpunktkarakteren i norsk fra ungdomsskolen utnyttes i undersøkelsen som grunnlag for en parameter for den enkelte elevs skriveferdighet; denne er altså egenrapportert. Elever med 6 eller 5 i norsk blir regnet for å ha "Sterke" skriveferdigheter, mens elever med 4 eller 3 er "Middels".

Skjema B var på to sider, men fordi dette skjemaet ble delt ut etter skrivingen, hadde jeg ikke like god kontroll på at alle elevene fylte ut dette skjemaet. 6 av de deltagende elevene fylte ikke ut skjema B. Informasjon fra Skjema B ble i liten grad utnyttet i undersøkelsen.

6.1.2 Utvalg av elever

Utvalget av elever ble gjort på grunnlag av tre ulike typer kriterier. For det første skulle utvalget nøytralisere mest mulig variasjon av mindre relevans for forskningsspørsmålene. Det innebar å utelukke elever uten norsk som morsmål (7), med nynorsk som hovedmål (ingen) og med standpunktkarakter i norsk lavere enn 3 (ingen). 4 elever ble dessuten utelukket fra studien fordi de hadde fritak fra håndskriving på skolen; jeg kjenner ikke begrunnelsen for fritaket, men antar at det som regel har sammenheng med enten dysleksi eller motoriske vansker.

For det andre skulle utvalget inneholde bare elever med et visst ferdighetsnivå i pc-bruk. Det innebar at bare elever som krysset av på "lett" for samtlige av tekstredigeringsaktivitetene i spørreskjema A, ble inkludert i studien. 17 elever ble utelukket av dette kriteriet. Dessuten måtte de ha egenrapportert egne tekstbehandlingsferdigheter som "bra" eller "ganske bra". 2

elever svarte "ganske dårlig" på dette spørsmålet og ble derfor utelukket, men én av disse hadde heller ikke norsk som morsmål. Spørsmålet i spørreskjemaet om hvilket verktøy eleven *foretrekker* å bruke til skriveoppgaver i norskfaget, ble ikke brukt som utvalgskriterium. Dette innebærer at 7 av elevene i det endelige utvalget rapporterer at det foretrukne verktøyet er "oftest hånd". Disse fordeler seg jevnt mellom å rapportere at de behersker tekstbehandlingsprogrammet "bra" og "ganske bra". De fordeler seg også jevnt over skriveferdighetsfaktoren, mens det er en viss overvekt av jenter, 5 mot 2. Dette innebærer at noen av elevene i utvalget kanskje har litt mer negative holdninger til eller litt lavere motivasjon for å bruke tekstbehandlingsverktøyet enn flertallet.

76 elever tilfredsstilte kriteriene over. Én av dem svarte ikke på spørsmålet om norskkarakter og ble derfor utelatt fra utvalget, én var syk og leverte bare én av oppgavene, og én leverte ved en feiltagelse to tastede tekster. De gjenværende 72 fordelte seg på parametrene kjønn og skriveferdigheter som vist i tabell 6-1 nedenfor.

Tabell 6-1: Elever som tilfredsstilte kriteriene, fordelt på kjønn og skriveferdighet

	gutter	**jenter**
middels	15	16
sterke	17	24

For det tredje ble spørreskjemainformasjonen brukt til å balansere utvalget, slik det går fram av tabell 6-2 nedenfor.

Tabell 6-2: Elever med to tekster i korpuset, fordelt etter kjønn og skriveferdighet

	gutter	**jenter**
middels	15	15
sterke	15	15

Informantutvalget som er brukt i studien, består av 60 elever. Blant de 72 elevene over ble det gjort et tilfeldig utvalg fra hver kombinasjon av kjønn og ferdighet slik at det gjenstod 15 elever i hver kategori. Utvalget er dermed *balansert* over de to parametrene kjønn og skriveferdighet, og altså *ikke tilfeldig* og *ikke representativt.* Begrunnelsen for å balansere utvalget på denne måten er å gi mer styrke til de parametriske hypotesetestene og modellene ved avvik fra premissene for testene (7.2.2), men denne gevinsten kommer med en viss kostnad, nemlig svekket representativitet og generaliserbarhet.

Det kan diskuteres om utvalget burde ha vært gjort bare blant de som rapporterer at de alltid eller oftest foretrekker å bruke pc. Men hvis jeg skulle ha brukt dette kriteriet sammen med kriteriet om ferdigheter i pc-bruk, måtte jeg enten ha redusert antall elever i utvalget eller ha utvidet rekrutteringsarbeidet. I lys av vanskelighetene ved å rekruttere informanter var det siste alternativet lite aktuelt. Jeg hadde på det tidspunktet avgjørelsen ble tatt, lite kunnskap om hvor store utvalg som ville være nødvendig for å oppnå signifikante resultater for de effektstørrelser som kunne være aktuelle, men i ettertid er det lite tvil om at det ville ha vært lite gunstig å redusere utvalgsstørrelsen, i lys av de relativt svake tendensene analysene har

avdekket. Derimot kunne det ha vært aktuelt å bruke et preferansekriterium for deltagelse *i stedet for* et ferdighetskriterium, men jeg valgte altså å beholde ferdighet som grunnleggende kriterium.

6.2 Tekster

Elevene skrev to argumenterende sakprosatekster. Alle elevene svarte på de to samme oppgavene. Oppgavene ble formulert på måter som i størst mulig grad skulle redusere forskjellene mellom de to skrivesituasjonene til et minimum, samtidig som oppgavene ikke skulle være så like at skrivingen fikk preg av å være en repetisjon med eventuelle uheldige øvings- eller slitasjeeffekter. De to oppgavene har fått kortnavnene A1 og A2 og er presentert i sin helhet i appendiks B8 og B9. Halvparten av elevene skrev A1 for hånd og A2 på tastatur; den andre halvparten gjorde det motsatt.

Oppgave A1: "Bøker eller data?"

- Tidsramme: 2 skoletimer
- Sjanger: leserinnlegg i avis
- Målform: bokmål
- Oppgavetekst: I et leserinnlegg om ungdommers medievaner ble følgende påstand fremsatt: "Gutter leser ikke bøker, men driver med data. Jenter driver ikke med data, men leser bøker".

  Skriv et leserinnlegg til en avis der du kommenterer og drøfter denne påstanden. Bruk overskriften "Bøker eller data?".

Oppgave A2: "Ungdomsfylla?"

- Tidsramme: 2 skoletimer
- Sjanger: leserinnlegg i avis
- Målform: bokmål
- Oppgavetekst: En MMI-undersøkelse viser at ungdommens drikkevaner har endret seg i negativ retning. I en lederartikkel i Hamar Arbeiderblad 6. november 2006 skriver redaktøren blant annet:

  "Hva gjør foreldrene når 14-15-åringene kommer og ber om øl til kveldens fest? Svaret bør være enkelt."

  Skriv et leserinnlegg der du først gjør greie for hva du tror kan være årsakene til den negative utviklingen, og deretter diskuterer synspunktet til redaktøren. Bruk overskriften "Ungdomsfylla?".

De to oppgavene har mange fellestrekk, nettopp med tanke på å redusere antall påvirkningsfaktorer på tekstene:

- Begge oppgavene innebar å skrive et leserinnlegg til en avis.
- Begge tekstene skulle formuleres som et tilsvar til tidligere aviskommentarer.
- Begge oppgavene gjengav et kort sitat som tekstene skulle ta utgangspunkt i.
- Begge disse sitatene var relatert til temaer som angår ungdommenes hverdag.
- Begge tekstene skulle skrives på bokmål, som er alle elevenes hovedmål.
- Begge oppgavene var formulert slik at elevene i liten grad ville ha nytte av Internett eller andre eksterne kilder.

Begge skriveøktene ble gjennomført på skolen innenfor to skoletimer, og øktene var en naturlig del av et undervisningsopplegg om argumenterende skriving som ble gjennomført i henhold til fagets årsplan. Tekstene ble levert inn til lærer for tilbakemelding og vurdering, og skrivingen fant dermed sted i en autentisk skolesituasjon. Alle elevene skrev oppgave A1 i den første skriveøkten og A2 i den andre skriveøkten. De to skriveøktene ble holdt med mellom 7 og 12 dagers mellomrom i de fire klassene. Elevene ble ikke orientert om oppgavenes tema på forhånd, men ettersom skriveøktene ikke ble holdt på samme dager i de fire klassene, kan jeg ikke utelukke at noen elever har snakket sammen om oppgavene. Det kan derfor være at elevene i noen klasser har vært bedre forberedt enn andre.

Opplegget medfører at teksten om "Bøker eller data" (A1) ble skrevet tidlig i undervisningsopplegget om argumenterende skriving, mens teksten om "Ungdomsfylla" (A2) ble skrevet etter at elevene hadde lært mer om argumenterende skriving, og antagelig for de fleste elevene etter at de hadde fått tilbakemelding på den første teksten. I en samtale mellom meg og lærerne etter skriveøktene rapporterte lærerne at de syntes oppgave A2 hadde vært mest vellykket, i den forstand at elevene i større grad hadde greid å argumentere i denne besvarelsen enn i besvarelsen til oppgave A1. At A2-tekstene i gjennomsnitt er lengre enn A1-tekstene, kan også tyde på at elevene fant det lettere å besvare denne oppgaven. Selv om oppgavene var formulert med tanke på å lage så like skrivesituasjoner som mulig, kan de sikkert ha appellert til elevene på forskjellig måte, men forskjellen i tekstlengde og systematiske språklige og tekstlige forskjeller mellom de to besvarelsene (se kapittel 9 – 11) *kan* også rett og slett skrive seg fra rekkefølgen de to tekstene er skrevet i. Rekkefølgen er dermed en mulig feilkilde for hele forsøket. I tilsvarende forsøk bør man forsøke å nøytralisere denne faktoren, men i dette forsøket var dette praktisk vanskelig ettersom alle elevene gikk på samme skole og man kunne regne med at de ville snakke med hverandre om oppgavene. Ved å gi den samme oppgaven først til alle elevene, ble risikoen for at noen av oppgavene ble kjent for noen av elevene før de skulle skrive dem, redusert.

Etter skriveøkten ble de håndskrevne tekstene fotokopiert av lærerne og overbrakt meg, mens de tastede tekstene ble sendt meg på epost.

6.3 Personvern

Prosjektet er meldt til Norsk Samfunnsvitenskapelige Datatjeneste (NSD), som fungerer som personvernombud for alle forskningsprosjekt som gjennomføres ved de statlige høgskolene, deriblant den daværende Høgskolen i Hedmark. (Se appendiks B1.)

Elevene ble orientert skriftlig om prosjektet gjennom et brev (appendiks B3) som ble delt ut av lærerne. Brevet inneholdt all relevant informasjon om anonymisering og informantenes rettigheter, i tillegg til at det ble opplyst at prosjektet "har som formål å undersøke *språket* i skolearbeid som elever har gjort på pc." (Utheving som i brevet.) Lærerne orienterte dessuten elevene muntlig i timen. Siden elevene var 16 år gamle, og prosjektet ikke hadde som formål å avdekke sensitive opplysninger, var det ifølge NSD ikke nødvendig å innhente foreldrenes tillatelse.

Elevene ble orientert om at det var frivillig å delta i prosjektet, og at de selv om de gav samtykke nå, kunne trekke seg fra prosjektet ved en senere anledning. Ingen av elevene har trukket seg fra prosjektet undervegs, men to av elevene gjennomførte ikke skrivingen som forutsatt, som beskrevet i 6.1.2 ovenfor. Dessuten var det en elev som ved en feil skrev begge oppgavene med samme skriveverktøy, og tekstene til denne eleven kunne derfor ikke benyttes i datamaterialet.

Elevene gav sin skriftlige tillatelse gjennom avkrysning og signatur på en svarslipp som ble samlet inn og håndtert av lærerne, men ikke videreformidlet til meg, slik at elevene kunne forbli fullstendig anonyme for meg.

NSD satte som forutsetning for godkjenning av prosjektet at alle tekster ble lest av lærerne *før* de ble videresendt til meg, slik at lærerne kunne holde tilbake eventuelle tekster som inneholdt sensitive opplysninger om navngitte eller lett identifiserbare personer. Ingen tekster ble holdt tilbake av en slik årsak, og det er etter min vurdering heller ikke noen tekster som inneholder slike opplysninger.

Ingen av tekstene inneholdt ikke-fiktive personnavn eller lett identifiserbare personer. Noen tekster inneholdt fiktive personnavn; disse er ikke endret. Noen tekster inneholdt stedsnavn. De tre lokale stedsnavnene som forekommer i tekstene, har jeg pseudonymisert (se 6.4.3.3.). Ikke-lokale stedsnavn som *Oslo* er ikke endret.

6.4 Korpusbygging

Dette delkapitlet omhandler korpusteknologien og tilpasning av tekstene og data fra spørreskjemaet til denne teknologien.

6.4.1 Overblikk

Dette avsnittet gir et overblikk over prosessen fra innsamling av tekster til ferdigbygd korpus klart for søking. Avsnittet gir kontekst til de etterfølgende delavsnittene, som tar for seg de enkelte trinnene i prosessen i detalj.

Håndskrevne tekster ble transskribert til elektronisk form ved at jeg skrev dem av i Word og lagret hver tekst som en fil med filnavn som indikerte oppgave og elev-id, men ikke skriveverktøy.

Alle tekster fikk korrigert ortografien slik at taggerprogrammet kan gjenkjenne det intenderte leksemet og bruke det som grunnlag for syntaktisk analyse. Ortografikorrigeringen ble gjort med XML-tagger som tar vare på informasjon om opprinnelig skrivemåte. Også feil i store forbokstaver ble korrigert på denne måten.

Alle tekster fikk korrigert tegnsetting slik at den i størst mulig grad rettledet taggerprogrammet mot riktig analyse. Tegnsettingen ble rettet i selve tekstene uten å opprettholde spor etter den opprinnelige tegnsettingen.

Tekstene ble på dette stadiet konvertert til XML-filer på tekstformat og flyttet fra Word til programmet Oxygen, som er et redigeringsprogram som er spesialtilpasset redigering av XML-filer (SyncRO Soft SRL, 2017).

Tekstene ble manuelt segmentert i t-enheter med XML-taggen `<t-unit>`. (Se definisjon av t-enhet i 4.1.2.) Fragmenter som ikke lot seg innpasse i noen t-enhet, fikk taggen `<frag>`. Overskrifter og underskrifter fikk også egne tagger som skiller dem fra brødteksten.

T-enhetene ble segmentert i subklaususer, som ble kategorisert og subkategorisert etter funksjonelle kriterier. Subklaususene ble markert med XML-taggen `<clause>`, med eventuell informasjon om kategori og subkategori notert som annoteringer i taggen. Taggene som segmenterer tekstene på denne måten, blir i korpussammenheng gjerne betegnet som *strukturelle attributter*.

Informasjon om elev og tekst ble lagt inn som egne XML-tagger i hodet av XML-filen.

XML-filene ble deretter importert til korpusløsningen. Som del av importen ble tekstene analysert av Oslo-Bergen-taggeren (Johannessen, [s.a.]), som først lemmatiserer teksten, deretter setter til morfologisk annotasjon og til slutt syntaktisk annotasjon.

Korpuset er annotert på en måte som gjør det mulig å søke etter fenomener på ulike språklige nivå, både ordform, lemmaform og automatiske annoteringer om morfologiske og syntaktiske egenskaper. Dessuten kan man søke etter strukturelle attributter, altså de annoteringer for syntaktiske segmenter som er registrert manuelt for t-enhet, subklausus og forskjellige typer av t-enheter og subklaususer. Søkekriterier kan også kombineres og slik danne både svært omfattende og svært spesialiserte søk.

Resultatlistene fra søk i korpusløsningen ble importert til MicroSoft Excel, der eventuell manuell bearbeiding av listene ble gjort. Dessuten ble trefflistene fra korpuset konvertert til en råskåreliste med antall treff for hver tekst.

Råskårelista i Excel ble deretter konvertert til en tekstfil og importert til statistikkprogrammet R (R Core Team, 2016), der de statistiske beregningene ble gjort.

6.4.2 Korpusteknologi

Dette prosjektet er hovedsakelig basert på korpusverktøyet Corpuscle (Meurer, 2012b, 2017), som i løpet av prosjektet har vært i kontinuerlig utvikling ved Uni Digital i Bergen. Utviklingen av Corpuscle begynte imidlertid ikke før prosjektet var godt i gang, og korpusbyggingen for elevtekstkorpuset baserte seg fra begynnelsen på korpusverktøyet for ASK – Norsk Andrespråkskorpus (Meurer, 2012a), delvis kombinert med søking med tekstsøkeverktøyet PowerGrep (Goyvaerts, 2009) i tekstfiler fra Oslo-Bergen-taggeren (Johannessen, Hagen, Lynum, & Nøklestad, 2012)[5]. Flere designvalg er tatt med tanke på den opprinnelige tekniske løsningen, og noen av disse valgene ville nok ha vært gjort annerledes dersom Corpuscle hadde lagt premissene fra starten.

I begynnelsen av prosjektet ble tekstene importert til en tilpasset versjon av korpussystemet for ASK, og da ble tekstene tagget med versjon 1 ("CG1" for *Constraint Grammar version 1*) av taggerprogrammet. Senere ble tekstene importert til en tilpasset versjon av Corpuscle, og da ble tekstene tagget med versjon 3 ("CG3") av taggerprogrammet. Det er selvfølgelig kvalitetsforskjeller mellom de to versjonene, men det er også noen mer prinsipielle forskjeller med konsekvenser som gjør at de er verdt å nevne:

- Lemmatiseringsalgoritmen er forskjellig i de to versjonene. CG3 benytter i større grad flerordsleksemer enn CG1. Siden data fra begge versjoner er med i analysene i denne avhandlingen, stemmer ikke tallene i resultatene alltid helt overens; forskjellene er imidlertid små.
- I CG1 ble all analyse disambiguert, slik at bare én analyse for hvert ord ble satt inn i korpuset. I CG3 står det alternative morfologiske og syntaktiske opplysninger for en del ord. 2650 ord har alternative lemmaformer. 9879 av 60239 ord har alternative morfologiske tagger.

6.4.3 Transskripsjon

De 60 håndskrevne tekstene har jeg transskribert til elektronisk form ved å skrive dem inn i Word. De 60 tastede tekstene har jeg også redigert slik at alle tekstene er i samme form. Dette underkapitlet presenterer metodene for denne transskripsjonen og redigeringen.

Transskripsjonen er motivert delvis ut fra den syntaktiske analysen og delvis av de teknologiske premissene.

---

[5] "Oslo-Bergen-taggeren er en robust morfologisk og syntaktisk tagger som er utviklet ved Universitetet i Oslo og Uni Computing i Bergen gjennom flere år. Taggeren består i dag av tre hovedmoduler: en preprosessor med sammensetningsanalysator og multitagger, en grammatikk-modul for morfologisk og syntaktisk disambiguering (constraint grammar) og en statistisk modul som fjerner siste rest av gjenstående morfologisk flertydighet (bare for bokmål). Grammatikk-modulen bruker en kompilator utviklet ved Syddansk universitet i Odense. Multitaggeren benytter fullformsleksikonet Norsk ordbank." (Tekstlaboratoriet, [s.a.])

Man kunne kanskje vente seg at elevenes håndskrift kunne skape utfordringer for transskripsjonen, men generelt var alle de håndskrevne tekstene lette å tyde. Bare én ordform var jeg i særlig tvil om tolkningen av. Dette ordet har jeg transskribert som \skrudlerier\, en nyskapning som jeg oppfatter som et synonym for "skriblerier".

6.4.3.1 Ortografi

Elevenes ortografi er rettet i tråd med taggerens ordliste, slik at analysen av leksem, morfologi og syntaks er basert på den intenderte ordformen. Også feil i store forbokstaver er rettet på denne måten, ettersom slike feil kan påvirke taggerens evne til å skille mellom proprier og appellativer. Alle ortografiske rettinger er gjort med en egen feil-tagg som markerer feilen og viser hvilken form eleven opprinnelig skrev.

(65) `<corr sic="desverre">dessverre</corr>`

Ettersom det er taggerens funksjon som er motivasjonen for å korrigere feil i ordformer, er det ikke nødvendig å korrigere ordformer som er oppført i taggerens ordliste selv om de ikke er i tråd med gjeldende rettskrivningsnorm. For eksempel korrigerte jeg ikke \gutta\ til \guttene\, ettersom ordformen ligger i taggerens ordliste som en bøyningsform av \GUTT\, selv om \gutta\ ikke er i henhold til gjeldende rettskrivning. Tilsvarende er det tilfeller der jeg har endret ordformer som er i tråd med rettskrivningen, men som forvirrer taggeren, for eksempel har jeg endret objektsforekomster av \de\ til \dem\ for å unngå at taggeren tagger disse forekomstene som subjekt, selv om man kan hevde at dette innebærer en endring av grammatiske forhold i teksten og slik kan påvirke studieobjektet.

Det er totalt 1153 ortografiske rettelser i korpuset, i den forståelse av ortografisk som jeg forklarer i de foregående avsnittene.

Syntaktiske avvik har jeg imidlertid ikke endret, ettersom det blant annet nettopp er aspekter ved elevenes syntaks jeg ønsker å belyse. Imidlertid er det en ikke triviell grenseoppgang mellom ortografi, morfologi og syntaks her; jeg har som nevnt for eksempel endret \de\ som objekt til \dem\, for å unngå at taggeren tolker pronomenet som subjekt og eventuelt snur om på klaususen. Det eneste syntaktiske avviket jeg *har* rettet, er at jeg har fjernet én forekomst av en ordform der eleven åpenbart har skrevet samme ordform to ganger ved en inkurie.

Det er helt nødvendig å endre ortografiske feil på den måten som er beskrevet, fordi feilstavede ordformer ofte fører til enten at taggeren knytter ordformen til feil leksem i ordlista, eller at taggeren tilegner ordformen taggen <ukjent> uten noen morfologisk informasjon. Feil eller manglende morfologisk informasjon om et ord har som potensiell konsekvens feil også i den syntaktiske taggingen, som i sin tur vil kunne påvirke analysen av andre ord og syntagmegrenser.

Det viktigste argumentet for å ledsage korrigeringene av feil-tagger, var at det gav muligheten for å gjøre oppfølgingsstudier av ortografiske forhold i tekstene. I det opprinnelige ASK-systemet var strukturelle attributter, herunder `<corr>`, behandlet teknisk på en måte som gjorde at det ikke hadde store konsekvenser for søk som ikke gjorde bruk av

strukturelle attributter. I Corpuscle, derimot, er strukturelle attributter behandlet teknisk som om de var ord, noe som innebærer at søk etter strenger der antall ord spiller en rolle, kan bli både ganske kompliserte å konstruere og ganske krevende for korpusmaskinen å utføre. I noen tilfeller har dette medført at aktuelle søk er blitt så tunge at de i praksis ikke har kunnet utføres. I lys av dette ville det ha vært en bedre løsning å lage to versjoner av korpuset, én med korrigerte ordformer uten feil-tagger og én med korrigerte ordformer *med* feil-tagger. I prosjektet i denne avhandlingen var det uansett ikke behov for feil-taggene, og det hadde vært gunstigere for gjennomføringen om de ikke var blitt satt inn i tekstene. I stedet burde ortografiske avvik bare ha blitt rettet direkte i tekstene før tekstene ble importert til korpusløsningen.

6.4.3.2 Tegnsetting

Jeg studerer ikke tegnsetting i denne avhandlingen, men når taggerprogrammet analyserer tekst, bruker det informasjon også fra tegnsettingen i analysen. For eksempel vil komma påvirke taggerens avgjørelse med hensyn til visse syntagmegrenser, særlig subklaususer.

Jeg har derfor også gjort endringer i tegnsettingen i tekstene, ikke for å gjøre tegnsettingen korrekt med hensyn til en viss standard, men med det for øye å hjelpe taggeren til rett analyse. For eksempel har jeg konsekvent satt inn komma etter subklaususer, og jeg har fjernet punktum eller erstattet dem med komma der punktumet opptrer midt i en t-enhet. Tegnsettingen har jeg endret fortløpende undervegs i transskriberingen og i korrekturen uten å markere endringene med egne feil-tagger. Dette betyr at korpuset ikke kan brukes til studier av tegnsetting. Det viktigste argumentet for å ikke feil-tagge tegnsettingsfeil, er at feil-taggene gjør søkestrengene mer kompliserte, og søkene tyngre. I lys av erfaringene med feil-taggene for ortografi må dette sies å ha vært et heldig valg.

Avsnitt er ikke del av analysen i denne studien, og jeg har derfor ikke lagt mye arbeid i transskriberingen av avsnittsmarkeringen i tekstene. Der avsnitt er tydelig markert i teksten, har jeg også markert avsnitt i transskripsjonen. Alle halvavsnitt, både i tastetekster og i håndtekster, har jeg enten fjernet eller gjort om til avsnitt, hovedsakelig basert på semantiske forhold i teksten. I håndtekstene er det ikke alltid lett å avgjøre om et halvavsnitt er intendert eller om det bare er et ordinært linjeskift. I enkelte tilfeller er en t-enhet delt mellom to avsnitt; i disse tilfellene har jeg prioritert å beholde den grammatiske formen og fjernet avsnittsmarkeringen.

6.4.3.3 Annet

Lokale stedsnavn ble pseudonymisert gjennom de fiktive stedsnavnene *Småby, Bygdeby* og *Tettsted*. Bare tre lokale stedsnavn er brukt i tekstene, og tilsvarende navn på aviser, skoler og andre lokale institusjoner er pseudonymisert på samme måte. Frekvensen av slike stedsnavn er så lav at disse endringene ikke har noen vesentlig innvirkning på de statistiske beregningene av leksikalske variabler.

I tekstene forekommer bare fiktive personnavn, så pseudonymisering av personnavn var ikke nødvendig., men en del av dem er plassert som signatur i teksten og er derfor ikke med i det tekstmaterialet som ble analysert for leksikosyntaktiske variabler. (Se 6.4.4.3 om signaturer og andre perifere deler av tekstene.)

6.4.4 Strukturell segmentering

Hele tekstkorpuset er manuelt segmentert i syntaktiske segmenter. Segmenteringen er gjort ved å sette inn XML-tagger (World Wide Web Consortium, 2016) som er spesialtilpasset analysene for dette prosjektet, og programmereren har tilpasset korpussystemene slik at systemet håndterer disse taggene slik den håndterer andre strukturelle attributter.

6.4.4.1 XML

XML-strukturen i tekstene er styrt av tre nivåer av regler. For det første er det generelle regler fra definisjonen av XML. Disse legger restriksjoner på den generelle strukturen i XML-elementene. Av de viktigste reglene er at XML-tagger opptrer parvis med en start-tagg og en stopp-tagg, som vist i eksempel (66). Stopp-taggen er identisk med start-taggen, men innledes av en skråstrek.

(66) `<t-unit>...</t-unit>`

En annen sentral regel styrer rekkefølgen av ulike XML-tagger. Forskjellige XML-tagger kan omslutte hverandre, som i (67), men element 2 må avsluttes før element 1. Eksempel (68) viser en struktur som bryter dette prinsippet og dermed er en ugyldig XML-struktur.

(67) `<t1>...<t2>...</t2>...</t1>`
(68) `<t1>...<t2>...</t1>...</t2>`

Eksempel (69) kan imidlertid være en gyldig struktur, men det er en forutsetning at det er de to ytre taggene som danner et par, og de to indre. Alle programmer som tolker XML-kode vil pare taggene i eksemplet på denne måten.

(69) `<t1>...<t1>...</t1>...</t1>`

For det andre styres strukturen av regler som gjelder en enkelt applikasjon. I elevtekstprosjektet har jeg konstruert et stilark (*stylesheet*) som for eksempel definerer taggenes egenskaper slik at tagger av typen `<clause>` bare kan opptre innenfor tagger av typen `<t-unit>`. Andre korpusløsninger trenger ikke ha denne begrensningen.

For det tredje har jeg brukt noen konvensjoner som ikke er definert hverken av XML eller av *stylesheet*. Først og fremst gjelder det at alle ord i løpeteksten skal være omsluttet av et strukturelt attributt, som regel `<t-unit>`.

Arbeidet med syntagmeannoteringen av tekstene ble gjort i programmet Oxygen XML Editor (SyncRO Soft SRL, 2017), som er et redigeringsprogram som er spesielt tilpasset

redigering av XML-filer. Programmet har automatisk feilsjekking og sikrer dermed mot formelle feil eller mangler i XML-strukturen.

De strukturelle attributtene har status som ord i korpuset i visse sammenhenger, men håndteres spesielt av korpusprogramvaren i andre sammenhenger.

6.4.4.2 Segmentering i syntaktiske segmenter

Tekstenes brødtekst ble først segmentert i t-enheter (definert i 4.1.2), slik eksempel (70) viser. Eventuelle overskrifter og underskrifter ble tagget med egne tagger og holdt utenfor den syntaktiske segmenteringen. I segmenteringen i t-enheter er prinsippet at t-enheten er en "maksimal" enhet; det vil si at mest mulig språklig materiale blir tilegnet hver t-enhet. Segmenter som ikke kan sies å tilhøre hverken foregående eller etterfølgende t-enhet, blir kategorisert enten som finitte fragmenter (71) eller som ekte fragmenter (72), som forklart i 4.1.3 ovenfor. Med tanke på senere analyser er det en fordel at de fragmenter som har en form for predikatskompleksitet og eventuelt subklausale strukturer, kan inkluderes i de analysene som retter seg mot slike egenskaper.

(70) <t-unit> Foreldre gir barna sine alkohol</t-unit> [A2-259]
(71) <t-unit type="frag"> fordi data er den nye verden.</t-unit> [A1-237]
(72) <frag> Tilbake til påstanden.</frag> [A1-260]

Taggene som representerer de strukturelle attributtene, blir plassert slik at eventuell tegnsetting alltid er plassert innenfor taggene.

Segmenter av typen `<t-unit>` kan subsegmenteres; finitte fragmenter tagges som en type t-enhet og kan dermed også subsegmenteres. Stilarket (gjengitt i C1) tillater imidlertid ikke t-enheter innenfor t-enheter, så en t-enhet kan kun subsegmenteres med subklaususer.

Ekte fragmenter tagges med `<frag>` og kan ikke subsegmenteres. De blir ikke behandlet videre, hverken av den manuelle segmenteringen eller av den automatiske taggingen. Corpuscle gir heller ikke treff på ord som står innenfor `<frag>`, mens det ASK-baserte systemet gjør det. Det innebærer at det rapporterte antall ord i tekstene avviker noe mellom de to systemene (se 6.4.2).

Subklaususer kan i likhet med t-enheter type-spesifiseres, og dette er den normale situasjonen for subklaususer. De aktuelle kategoriene er de tre funksjonelle subklausustypene, i tillegg til en egen kategori for underklaususer på hovedklaususform. Mange relativklaususer og de fleste adverbialklaususer er dessuten subkategorisert som forklart i 4.1.4, men denne subkategoriseringen har ingen praktisk betydning i de analysene som er presentert i avhandlingen. (73) illustrerer noen av de vanligste taggene i et litt mer komplekst eksempel.

(73) <t-unit>Og ja, for mange prektige og staselige foreldre <clause type="relativ">som "aldri" gjorde noe galt <clause type="adverbial" fn="tid">da de var unge,</clause></clause> er det sikkert det.</t-unit> [A2-312]

I tabell 6-3 nedenfor er de mest relevante attributtkodene gjengitt; i tillegg finnes det enkelte mer tekniske koder som ikke kommer frem i avhandlingsteksten eller korpussøkene.

Tabell 6-3: Strukturelle attributter. De viktigste XML-kodene som er brukt som strukturelle attributter i korpuset.

`<t-unit>`	T-enhet
`<t-unit type="frag">`	Fragment med finitte elementer
`<clause>`	Subklausus. Hovedklaususer er ikke eksplisitt notert.
`<clause type="adverbial">`	Adverbial subklausus
`<clause type="helsetning">`	Subklausus på hovedklaususform[6]
`<frag>`	Setningsfragment
`<corr>`	Ordform med korrigert rettskrivning
`<head>`	*Heading,* overskrift
`<closing>`	Underskrift eller ikke-klausalt avslutningsformular
`<p>`	*Paragraph*, avsnitt
`</clause>`	Slutt på subklausus
`</t-unit>`	Slutt på t-enhet

6.4.4.3 Annet

De fleste tekstene har en overskrift. Overskriften er markert med en egen tagg og er ikke gjenstand for leksikalsk eller syntaktisk analyse av taggerprogrammet. Noen av tekstene har også en underskrift, noen ganger med en slags hilsen eller annen type formulering som ikke inneholder finitte elementer. Disse er også markert med en egen tagg og er heller ikke gjenstand for analyse av taggerprogrammet.

Avsnitt er markert med en egen tagg, men denne markeringen har ingen praktisk betydning i prosjektet.

6.4.4.4 Korrektur av manuelt registrerte data

Jeg har selv lest korrektur på transskripsjonene av håndtekstene umiddelbart etter transskripsjon. Det har ikke på senere tidspunkt vært lest systematisk korrektur av transskripsjon, ortografisk korrigering, tegnsetting eller strukturelle attributter, men alle nivåer av transskripsjonene er blitt korrigert etter hvert som enkeltfeil er blitt avdekket gjennom analysene. Etter at korpuset ble flyttet fra det ASK-baserte systemet til Corpuscle, ble prosedyren for oppdatering av tekstmaterialet vanskeligere, og feil som er blitt avdekket etter at den nye teknologiske løsningen ble tatt i bruk, er ikke endret i korpuset.

[6] Betegnelsen 'helsetning' i taggen ble valgt i en fase av prosjektet da teorigrunnlaget og terminologien ikke var bestemt. Det var forbundet med uforholdsmessig mye arbeid å endre taggnavnet så det skulle være i samsvar med terminologien som blir brukt i prosjektet, og det har ingen praktiske konsekvenser.

Avgjørelsen om ikke å gjennomføre en systematisk korrektur ble tatt på grunnlag av en vurdering av tidsressurser. Det er mitt inntrykk gjennom arbeidet med materialet at feilprosenten på de forskjellige nivåene er ganske lav, og at den er vesentlig lavere enn feilprosenten fra taggeren. En bedring av nøyaktigheten i transskripsjon, ortografi, tegnsetting og segmenttagging ville i så fall ikke hatt avgjørende innvirkning på den totale nøyaktigheten i systemet. Jeg har imidlertid ikke gjort spesifikke beregninger av nøyaktigheten hverken i den manuelle transskripsjon eller annotering eller i den automatiske annotering.

6.4.5 Automatisk tagging

Etter den manuelle taggingen med strukturelle attributter ble tekstene analysert av Oslo-Bergen-taggeren (Tekstlaboratoriet, [s.a.]), som parset teksten i ord og tilegnet hvert ord en lemmaform, ordklasse, informasjon om bøyningskategorier og syntaktisk funksjon.

Kombinasjonen av de manuelle strukturelle attributtene og de automatiske leksikogrammatiske attributtene er det som gir denne korpusløsningen den søkefunksjonaliteten som er nødvendig for å besvare forskningsspørsmålene i prosjektet.

7 Statistisk analyse

Dette kapitlet presenterer de viktigste statistiske metodene som blir brukt i avhandlingen, og drøfter noen av dem. De allment kjente prinsippene for statistiske metoder som er omtalt, bygger på generelle lærebøker (Baayen, 2008; Crawley, 2005, 2007; Gries, 2009) der ikke andre kilder er nevnt. Prinsippene for prinsipalkomponentanalyse blir ikke behandlet her, men blir presentert i begynnelsen av kapittel 12.

Kapitlet begynner med en kort presentasjon av verktøyene som er brukt (7.1), etterfulgt av en gjennomgang av prinsippene for statistisk hypotesetesting (7.2), som dessuten inneholder et avsnitt om testpremisser (7.2.2) som også er relevant for det påfølgende delkapitlet om statistiske modeller (7.3). Til slutt i kapitlet kommer et kort delkapittel om målestokk for språklige variabler (7.4) og et kort delkapittel som tar opp noen momenter i forbindelse med beregning av korrelasjoner med tekstlengde (7.4.2).

7.1 Statistikkprogrammet R

Alle de statistiske analysene i denne avhandlingen er utført i statistikkprogrammet R (R Core Team, 2016), som egentlig er et programmeringsspråk eller programutviklingsmiljø som er særlig godt egnet til statistiske beregninger. De fleste analysene og diagrammene er utført i versjon 3.2.4 av R, men noen av analysene er gjort i tidligere versjoner. I tillegg til standardinstallasjonen av R har jeg brukt R-programpakkene `languageR` (Baayen, 2008), `gvlma` (Peña & Slate, 2014) og `lattice` (Sarkar, 2008). Jeg har også skrevet enkelte R-funksjoner selv, med formler for utregning av visse parametre som standardfunksjonene ikke leverer direkte. Også disse presenterer jeg i sin sammenheng undervegs, og koden er dessuten gjengitt i appendiks E.

En del av de statistiske analysene i denne avhandlingen blir dokumentert gjennom tekstutskrifter fra R, og i disse tilfellene blir resultatene fra R gjengitt akkurat som de fremkommer i programmet, med engelsk tekst og med engelsk desimalpunktum. I de fleste diagrammene er tekst og tall gjengitt på norsk og med norsk desimalkomma, men i enkelte av diagrammene har det vært vanskelig å gjennomføre en slik konvertering; i slike tilfeller forekommer det også engelsk tekst og desimalpunktum i diagrammene.

Til redigering av R-programkode har jeg brukt redigeringsprogrammet Tinn-R (Faria, Grosjean, Jelihovschi, & Farias, 2015), som er spesielt tilpasset dette formålet.

7.2 Statistisk hypotesetesting

Underkapitlet presenterer kort de viktigste statistiske hypotesetestene som brukes i avhandlingen, inkludert formålet med testen, premissene for bruk og fordeler og ulemper ved de enkelte testene.

7.2.1 Generelt om hypotesetesting

Statistisk hypotesetesting dreier seg om å undersøke utvalg fra en eller flere populasjoner med tanke på å avdekke egenskaper ved populasjonen eller populasjonene. I denne avhandlingen som i mange andre forskningsprosjekter innen humaniora dreier det seg om en populasjon av personer. Utgangspunktet er at vi har en hypotese om en tendens i populasjonen, typisk at det finnes en forskjell i en variabel mellom to grupper av individer, eller at det finnes en sammenheng mellom to variabler målt på de samme individene. Vi måler variablene i utvalget av individer og vurderer om tendensen er så sterk i utvalget at en tilsvarende tendens trolig også finnes i populasjonen, eller om tendensen ikke er så sterk, men heller et utslag av tilfeldig variasjon.

Slik generalisering fra egenskaper i utvalget til egenskaper i populasjonen forutsetter at utvalget er tilfeldig og representativt. Et tilfeldig utvalg av elever forutsetter en trekking av elever der hver elev i populasjonen har like stor sannsynlighet for å bli med i utvalget. En tilfeldig trekking i denne forstand har vi sjelden anledning til å foreta i skoleforskning; snarere er det vanlig at utvalget kommer fra én eller noen få skoler, slik det gjør i denne undersøkelsen. I tillegg til at utvalget ikke er tilfeldig, er det heller ikke representativt, ettersom det er balansert på kjønn og ferdighet. Begge disse egenskapene peker mot at man skal være litt mer forsiktig med å konkludere om generalisering til populasjonen.

Mer formelt består hypotesetestingen i at man vurderer en sannsynlighet knyttet til en nullhypotese, der nullhypotesen består i at det ikke finnes en slik tendens i populasjonen som hypotesen postulerer, altså for eksempel en forskjell mellom to grupper i en variabel. Dersom det finnes en forskjell i utvalget, regner en statistisk test ut sannsynligheten for at en forskjell av den størrelsen eller større kunne finnes i et tilfeldig utvalg dersom nullhypotesen var sann. Denne sannsynlighetsverdien kalles gjerne testens p-verdi. Dersom p-verdien er under en viss grense, som vi kaller signifikansnivået, ofte denotert α, regner vi det som en støtte til eller bekreftelse av hypotesen. Dersom sannsynligheten derimot er over signifikansnivået, regner vi det som en bekreftelse av nullhypotesen. I så fall kan man si at hypotesen er tilbakevist, eller – litt mer forsiktig – at testen ikke har kunnet bekrefte eller underbygge hypotesen.

Dersom hypotesetesten gir støtte til hypotesen, kaller vi resultatet *signifikant*; i motsatt fall er det ikke-signifikant. Det er en vanlig misforståelse at signifikans betyr at tendensen er sterk eller viktig på en eller annen måte, men signifikans betegner altså kun den helt tekniske egenskapen at sannsynlighetsverdien fra hypotesetesten er under et visst nivå. Det er en sammenheng mellom tendensens styrke, utvalgets størrelse og testens p-verdi, men tendensens styrke må vurderes ut fra egne analyser og egne mål, som gjerne kalles effektmål eller effektstørrelse. Tendensens viktighet må derimot vurderes ved tolkning av resultatene i den kontekst de er relevante for undersøkelsens formål. Det er også en vanlig misforståelse at p-verdien betegner sannsynligheten for at det finnes en tendens i populasjonen; p-verdien betegner bare sannsynligheten for at et tilfeldig utvalg med de aktuelle egenskaper trekkes fra en populasjon der nullhypotesen er sann, og disse to sannsynlighetene er *ikke* like.

I forbindelse med statistisk hypotesetesting på utvalg fra populasjoner snakker vi alltid om sannsynlighet; vi kan aldri slå fast noe med absolutt sikkerhet så lenge vi ikke måler på hele populasjonen, noe vi sjelden har anledning til å gjøre. Det betyr at vi aldri er *helt* sikre på om en tendens gjelder for en populasjon; vi er bare mer eller mindre usikre. Usikkerheten innebærer at konklusjonen kan være feil, og en slik feil kan være av to typer. Hvis vi tilbakeviser nullhypotesen selv om den faktisk gjelder i populasjonen, har vi et falskt positivt resultat, noe som gjerne blir kalt for en type-1-feil. Den motsatte situasjon, at vi beholder nullhypotesen selv om den faktisk ikke gjelder i populasjon, innebærer et falskt negativt resultat, kalt type-2-feil. Valget av signifikansnivå, α, innebærer å velge en viss balanse mellom risikoen for de to feiltypene. Dersom vi ønsker å senke risikoen for å begå type-1-feil, må vi velge en lav verdi for α, men dette medfører en økt risiko for type-2-feil. Høyere verdier av α reduserer risikoen for type-2-feil og øker risikoen for type-1-feil. Et vanlig valg for α i humaniora er 0,05, som gjerne vurderes som en fornuftig balanse for forskningsspørsmål som er typiske for humanistiske vitenskaper, og det er denne verdien av α jeg velger for de statistiske analysene i denne avhandlingen. $\alpha = 0{,}05$ innebærer at vi aksepterer en risiko på 5% for å rapportere et falskt positivt resultat. Imidlertid er det viktig å være bevisst at den valgte α-verdien faktisk er et relativt tilfeldig valgt tall, og at det ikke er noen vesentlig forskjell på utvalg som gir $p = 0{,}049$, og utvalg som gir $p = 0{,}051$, selv om det ene resultatet betegnes som signifikant og det andre ikke. Det er likevel vanlig å betrakte α som en absolutt grense og signifikans som en dikotom størrelse; jeg mener det kan være fornuftig også å se på p-verdiens størrelse og legge den til grunn for en vurdering av grad av usikkerhet. Jeg oppgir derfor ofte avrundede p-verdier i denne avhandlingen.

7.2.2 Premisser og premisstesting

Mange statistiske hypotesetester forutsetter visse egenskaper ved dataene for at testresultatet skal være gyldig. For å bruke slike tester må man sannsynliggjøre at premissene er oppfylt, for eksempel med egne tester eller inspeksjon av dataene gjennom diagrammer. Dersom premissene ikke er oppfylt, forårsaker det at man endrer risikoen for både type-1- og type-2-feil, uten at man nødvendigvis vet hvilken retning endringen går i.

Fremstillingen i dette underkapitlet forutsetter en viss kjennskap til t-test og anova, men disse blir ikke presentert og forklart før i henholdsvis 7.2.3.1 og 7.3.2.

7.2.2.1 Uavhengige observasjoner

Et svært viktig og ufravikelig premiss er prinsippet om uavhengige observasjoner. Dette er et allment prinsipp som gjelder for alle statistiske tester som er brukt i denne avhandlingen, og består i at hver observasjon i materialet som skal testes, ikke er påvirket av en annen observasjon i materialet. I praksis innebærer dette for korpusstudier først og fremst at hver observasjon skriver seg fra én og bare én informant. I prinsippet kunne man hevde at observasjoner som skriver seg fra elever som har hatt den samme læreren, heller ikke er helt uavhengige, men jeg mener at dette ville være en altfor streng forståelse av uavhengighet, som i praksis ville umuliggjøre mange typer forskning på elever. Vi skal imidlertid være oppmerksom på at elevene alle går på samme skole, og at generaliseringer av resultatene til

hele Norges populasjon av elever i VG1 med studiespesialisering må gjøres med en viss forsiktighet.

Dessverre er premisset om uavhengige observasjoner ofte brutt i korpuslingvistiske studier, slik Levshina påpeker:

> It may happen that several occurrences of the preposition come from a text written by the same author. In that case, the corresponding observations would be dependent. However, this assumption is often relaxed in corpus linguistics [...]. (Levshina, 2015, s. 212)

At dette er vanlig, gjør det imidlertid ikke mer gyldig, og mange kvantitative studier rapporterer resultater som er ugyldige på grunn av denne metodologiske feilen, ofte ved at forskeren bruker kjikvadrat-test og lar hver enkelt forekomst av et fenomen telle som en observasjon i stedet for for eksempel å regne ut gjennomsnittsverdier for hver informant og bruke disse gjennomsnittsverdiene som observasjoner i hypotesetesten.

> Chi-square is a much-abused test in second language research studies, and often one of its assumptions (that of independence of data) is violated as a matter of course. (Larson-Hall, 2010, s. 206)

> It is not uncommon to find cases in which the assumption of independence of observations is violated [in chi-square tests], usually by having the same participant respond more than once. [...] This kind of error is easy to make, but is an error nevertheless. The best guard against it is to make certain that the total of all observations (N) equals precisely the number of participants in the experiment. (Howell, 2007, s. 152, 153)

I denne undersøkelsen er antall informanter N = 60, og antall uavhengige observasjoner er derfor maksimalt 60. I noen sammenhenger bruker jeg Pearsons korrelasjonstest på alle 120 tekstene, til tross for at de er skrevet av bare 60 elever. I disse tilfellene bruker jeg imidlertid ikke testen som hypotesetest, men bare som et verktøy for å kvantifisere styrken på korrelasjonen mellom to variabler. Jeg peker i disse tilfellene dessuten på at ikke alle observasjonene er uavhengige.

7.2.2.2 Normalitet

De hypotesetester som kalles parametriske tester, forutsetter at populasjonen har en kjent distribusjon, slik at testen kan benytte den teoretiske distribusjonens *parametre* i utregningene. For de parametriske testene som er brukt i denne avhandlingen, først og fremst t-test (7.2.3.1), Pearsons korrelasjonstest (7.2.4) og anova (7.3.2.1), er det *normaldistribusjon* i populasjonen som er forutsetningen – eller mer presist én av forutsetningene.

Det er verdt å merke seg at forutsetningen dreier seg om *populasjonens* distribusjon og ikke utvalgets. I vel utforskede forskningsfelt har man gjerne mye kunnskap om distribusjoner i populasjonene, og man kan kanskje uten videre gå ut fra at de er normalfordelt. Innenfor lingvistikk finnes det for eksempel for leksikalske tekstvariabler en del slik kunnskap om distribusjon (Baayen, 2001), men mange syntaktiske forhold er mindre utforsket. Det er derfor som regel nødvendig å analysere *utvalget* med tanke på om det er sannsynlig at populasjonen som utvalget er tatt fra, er normalfordelt. Til dette formålet kan man blant

annet benytte statistiske hypotesetester for normalitet, som på samme måte som hypotesetester for andre egenskaper vurderer en populasjons normalitet ut fra egenskapene til et utvalg fra populasjon

En vanlig metode for testing av normalitet er Shapiro-Wilks normalitetstest (Gries, 2009, s. 150; Larson-Hall, 2010, s. 84), som er en hypotesetest for normalitet som først ble publisert av Shapiro og Wilk (Shapiro & Wilk, 1965) (*NIST/SEMATECH e-Handbook of Statistical Methods*). Testen resulterer i en W-verdi med tilhørende p-verdi, som vist i eksemplet fra R under.

(74)
```
> shapiro.test(pos$substF[Tast])
W = 0.97901, p-value = 0.3882
```

Eksemplet viser Shapiro-Wilk-testen på variabelen substantivfrekvens i tastetekstutvalget. I dette tilfellet forteller p-verdien at dersom utvalg av denne størrelsen (N = 60) ble trukket fra en normalfordelt populasjon, ville 39 % av utvalgene avvike like mye eller mer fra normalfordeling som variabelen `pos$substF[Tast]` (substantivfrekvens i tastetekstene) gjør. På grunnlag av dette er det vanlig å trekke konklusjonen at `pos$substF[Tast]` sannsynligvis stammer fra en normalfordelt populasjon, og at parametriske tester som fordrer normalfordeling, kan benyttes på utvalget.

Det ser ut til å være vanlig å benytte $\alpha = 0{,}05$ som grenseverdi (signifikansnivå) for Shapiro-Wilks normalitetstest, på linje med hva som er vanlig i annen statistisk hypotesetesting innenfor områder som ikke er godt utforsket (Crawley, 2007, s. 281-282), men få innføringsbøker drøfter denne praksisen. Crawley (s. 281) omtaler Shapiro-Wilks normalitetstest, og han nevner bare denne ene normalitetstesten, men han sier også kort at et kvantil-kvantil-diagram ofte er den beste normalitetstesten. Crawley (s. 282) nevner $p < 0{,}05$ i det avsnittet han omtaler Shapiro-Wilks normalitetstest, men han knytter ikke denne grenseverdien eksplisitt til normalitetstesten. Baayen (2008, s. 73 og s. 76) bruker testen uten å kommentere α. Gries (2009, s. 150) bruker $\alpha = 0{,}05$ uten å drøfte verdien. Larson-Hall (2010, s. 84) bruker $\alpha = 0{,}05$, men peker også på at normalitetstester "often suffer from low power" (s.84), og argumenterer for bruk av visuelle metoder, men drøfter ikke valg av α i seg selv. Lowie og Seton (2013, s. 92) bruker $\alpha = 0{,}05$ uten å drøfte verdien, men anbefaler visuell inspeksjon ved siden av, særlig for store utvalg.

Logikken i en hypotesetest for normalitet er den samme som logikken i andre hypotesetester. Nullhypotesen for Shapiro-Wilk-testen, og andre normalitetstester, er at populasjonen er normalfordelt. Dersom $p > \alpha$, bekreftes nullhypotesen, og populasjonen regnes for normaldelt; dersom $p < \alpha$, tilbakevises nullhypotesen og populasjonen regnes for ikke-normalfordelt. I det siste tilfellet regner man med at parametriske tester som forutsetter normalitet, ikke kan benyttes på utvalget på en valid måte.

Formålet med normalitetstesting er imidlertid på sett og vis det motsatte i forhold til formålet med vanlig null-hypotesetesting. Et ikke-signifikant ($p > \alpha$) resultat fra normalitetstesten innebærer dermed et *positivt* funn i den forstand at funnet legitimerer bruk av en parametrisk test på variabelen. Det vil si at selv om normalitetstesten rapporterer at utvalgene vil avvike

tilsvarende mye fra normalitet i bare 8 % av tilfeldige utvalg fra populasjonen, slik tilfellet er for variabelen `pronF` (pronomenfrekvens) i håndtekstene i (75), regner man det som sannsynlig at utvalget stammer fra en normalfordelt populasjon.

(75)
```
> shapiro.test(pos$pronF[Hånd])
        Shapiro-Wilk normality test

data:  pos$pronF[Hånd]
W = 0,9643, p-value = 0,07637
```

Jeg vil knytte to kommentarer til denne praksisen. For det første virker $\alpha = 0{,}05$ intuitivt som en lite konservativ grenseverdi i denne sammenhengen. I ordinær nullhypotesetesting er grenseverdien satt så lavt som 0,05 for å redusere faren for å begå type-1-feil, altså å avvise nullhypotesen i de tilfeller nullhypotesen er sann. Vi ønsker å være ganske sikre på at nullhypotesen ikke gjelder i populasjonen, før vi rapporterer om positive funn. I 5 % av utvalg i de situasjoner der nullhypotesen er sann, vil det likevel være slik at vi rapporterer positive funn, og dette er normalt en risiko for feilrapportering vi mener er akseptabel. Hvis vi ønsker å være enda sikrere på ikke å begå type-1-feil, senker vi α, for eksempel til 0,01.

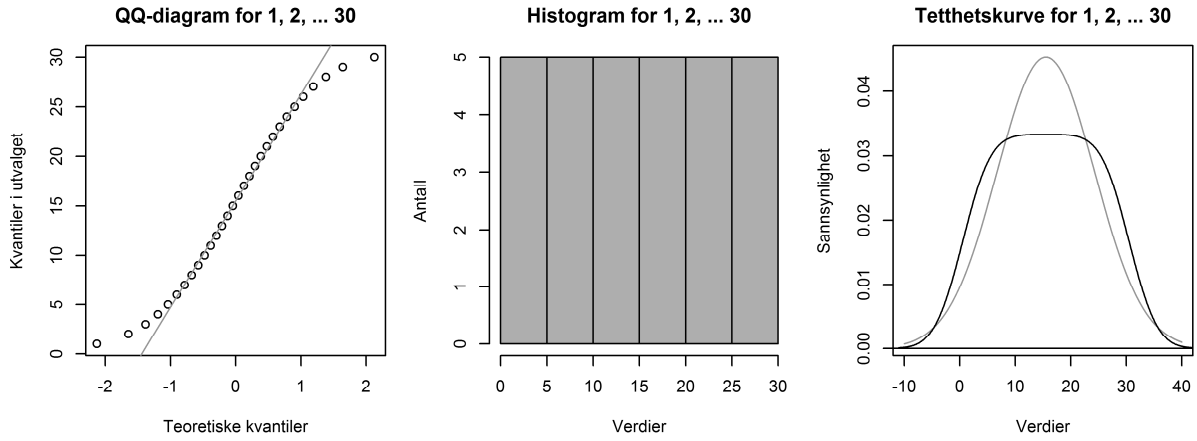
I normalitetstesting er det som nevnt over vanligvis slik at $p > \alpha$ er gunstig for analysen fordi dette legitimerer bruk av en parametrisk test (Gries, 2009, s. 50), og vi ønsker å bruke en parametrisk test når det er mulig, fordi parametriske tester normalt har større styrke enn ikke-parametriske tester til å avdekke reelle tendenser i populasjonen. Men dersom vi begår en type-2-feil i normalitetstesten og konkluderer med at utvalget kommer fra en normalfordelt populasjon når populasjonen faktisk ikke er normalfordelt, vil den parametriske nullhypotesetesten kunne gi ugyldige resultater. Det virker derfor som en mer konservativ tilnærming å *heve* α for å redusere faren for type-2-feil i normalitetstesten.

For det andre er det viktig å merke seg at p-verdien faktisk ikke angir sannsynligheten for at utvalget stammer fra en normalfordelt populasjon, selv om praktisk ordbruk ofte antyder nettopp dette, for eksempel:

> The Shapiro-Wilk test, proposed in 1965, calculates a W statistic that tests whether a random sample, x_1, x_2, ..., x_n comes from (specifically) a normal distribution. (*NIST/SEMATECH e-Handbook of Statistical Methods*, 2012, 7.2.1.3)

Det testen derimot rapporterer, er altså sannsynligheten for at et utvalg med slike egenskaper skulle bli trukket dersom populasjonen er normalfordelt. Dette er ikke det samme, blant annet fordi vi gjerne ikke kjenner populasjonens naturlige egenskaper i utgangspunktet. Shapiro-Wilk-testen vil gjerne rapportere høye p-verdier for små utvalg som det kan være mer sannsynlig å regne med har kommet fra populasjoner med andre distribusjoner. Som supplement til Shapiro-Wilk-testen bør man derfor inspisere utvalgenes distribusjonelle egenskaper visuelt, slik jeg nevnte at Lowie og Seton, Crawley og Larson-Hall peker på. Nyttige diagramtyper er kvantil-kvantil-diagram, histogram og tetthetskurver. Diagrammene i figur 7-1 nedenfor viser disse tre diagramtypene for den klart uniformfordelte distribusjonen av heltallene fra 1 til 30. Alle diagrammene viser at fordelingen trolig ikke

skriver seg fra en normalfordelt populasjon, men Shapiro-Wilk-testen går god for at et slikt utvalg kan stamme fra en normalfordelt populasjon, $W \approx 0{,}958$, $p \approx 0{,}28$.



Figur 7-1: Diagrammer som visualiserer (mangelen på) normalitet i tallrekken 1..30.

Det viser seg at testen virker ganske dårlig for små og middels store uniformfordelte utvalg. Dette er i tråd med Shapiro og Wilks egne vurderinger i 1965, der de (s. 608) sier at "W-testen" er særlig følsom for avvik som skyldes skjevhet eller lange haler. Deres empiriske styrkeberegninger gir vesentlig svakere resultater for uniform fordeling, som kan regnes som en form for kurtoseavvik, enn for lognormal fordeling (N = 20), som innebærer skjevhet.

Når det gjelder alternativer til Shapiro-Wilk-testen, er Kolmogorov-Smirnovs ettutvalgstest mye brukt. Howell (2007, s. 79) advarer mot denne testen for å teste normalitet og siterer D'Agostino & Stephens (1986): "The Kolmogorov-Smirnov test is only a historical curiosity. It should never be used."[7]. Likevel nevnes den som et alternativ i mange lærebøker (Baayen, 2008, s. 73; Crawley, 2007, s. 316; Gries, 2009, s. 151; Larson-Hall, 2010, s. 74 med et visst forbehold; Lowie & Seton, 2013, s. 47 med et visst forbehold) og den tilbys uten forbehold i en vanlig statistikkpakke som SPSS (Lowie & Seton, 2013, s. 140), riktignok med Lilliefors-korrigering. Også Ghasemi and Zahediasl (2012, s. 487, 489) advarer mot å teste normalitet med Kolmogorov-Smirnov-testen, også med Lilliefors-korrigering, på grunn av mangel på styrke. Mangel på styrke vil i dette tilfellet innebære at bruk av testen vil lisensiere parametrisk testing i tilfeller der det ikke er forsvarlig. Ghasemi og Zahediasl anbefaler å bruke Shapiro-Wilk-testen.

Til tross for problemene som er beskrevet over, er det likevel to momenter som taler for nytten av Shapiro-Wilk-testen i denne avhandlingen. For det første er det mye som taler for at lognormale eller andre typer skjeve fordelinger er mer naturlig forekommende enn uniforme eller andre kurtose-avvikende fordelinger i data av den typen vi studerer, og for det andre har t-testen tilnærmet samme gyldighet for uniforme data som for normalfordelte

[7] D'Agostino, R. B. & Stephens, M. A. (1986). Goodness-of-fit techniques. New York: Marcel Dekker.

(7.2.3.1), mens den fungerer en del dårligere for lognormale data. Shapiro-Wilk-testens manglende følsomhet for uniforme data har dermed liten eller ingen praktisk betydning.

I denne avhandlingen blir stort sett Shapiro-Wilks normalitetstest benyttet, og jeg har også valgt å følge vanlig praksis ved å benytte $\alpha = 0{,}05$ som grenseverdi, men dersom p-verdien nærmer seg α, kan det være grunn til å sammenligne resultatene fra parametriske tester med resultater fra ikke-parametriske tester.

Ved toutvalgstester som for eksempel t-testen er kriteriet at *begge* utvalgene kommer fra normalfordelte populasjoner; om man for eksempel deler variabelen for substantivfrekvens i to etter parameteren kjønn, må både guttetekstutvalget og jentetekstutvalget testes for normalitet. Dersom det er forskjellen mellom for eksempel tastetekstverdiene og håndtekstverdiene for samme elev som skal testes, er det differanseverdienes normalitet som er relevant.

7.2.2.3 Varians

I litteraturen blir ofte lik varians i begge utvalgene nevnt som et premiss for t-testen (Field, Miles, & Field, 2012, s. 373). Lik varians kan testes for eksempel med F-testen dersom utvalgene skriver seg fra normalfordelte populasjoner (Baayen, 2008, s. 81-82), mens en ikke-parametrisk test som for eksempel Fligner-Killeens test (Gries, 2009, s. 220) kan brukes dersom utvalgene ikke er normalfordelte.

Imidlertid anbefaler Ruxton (2006) ikke testing av dette premisset. Ruxton viser at Welch' versjon av t-testen, som korrigerer for ulik varians, er bedre å bruke i situasjoner med ulik varians, og like god som den opprinnelige Students t-test i situasjoner med lik varians. Welch' t-test er også standardvalget for t-test i R, og det er denne versjonen av t-testen som er brukt i denne avhandlingen, dersom ikke annet er sagt. For parede t-tester er ikke lik varians direkte relevant.

7.2.2.4 Premisstesting for variansanalyse (gvlma)

I anova-modellering (7.3.2.1) testes premissene på modellene som blir bygget gjennom analysen, og testene foregår derfor i etterkant av analysen, i motsetning til for eksempel ved t-test, der premissene helst testes i forkant av analysen.

Det finnes egne verktøy for premisstesting for anova, og jeg bruker først og fremst funksjonen `gvlma` fra R-biblioteket med samme navn (Peña & Slate, 2014), beskrevet av Peña og Slate (2006). `Gvlma` tester fire premisser for anova, som vist i deler av utskriften fra anvendelsen av `gvlma` for variabelen gjennomsnittlig ordlengde (hentet fra 9.1.3):

```
gvlma(x = lm(lexD$ordlengde ~ kjønn))

                     Value p-value                Decision
Global Stat         3.005e+00  0.5570 Assumptions acceptable.
Skewness            2.460e+00  0.1168 Assumptions acceptable.
```

```
Kurtosis            5.407e-01  0.4621 Assumptions acceptable.
Link Function      -7.125e-15  1.0000 Assumptions acceptable.
Heteroscedasticity  4.331e-03  0.9475 Assumptions acceptable.
```

De to premissene skjevhet (*skewness*) og kurtose (*kurtosis*) er knyttet til to forskjellige typer avvik fra normalitet. Skjevhet dreier seg om hvorvidt distribusjonen er symmetrisk om middelverdien, mens kurtose dreier seg om hvorvidt distribusjonen har den grad av spisshet som normalfordelingen har. Skjevhet kan være med hale mot venstre (negativ) eller med hale mot høyre (positiv), mens kurtoseavviket kan dreie seg om for spiss (leptokurtisk) eller for butt (platykurtisk) distribusjonstopp. En leptokurtisk distribusjon innebærer en økning av sannsynligheten for ekstreme verdier i forhold til normalfordelingen. *Link function* viser hvorvidt distribusjonen har avvik fra linearitet, mens heteroskedastisitet (*heteroscedasticity*) er et uttrykk for at variansen (eller variansen i residualene) ikke er uniform gjennom distribusjonen (Crawley, 2007, s. 340). Et vanlig problem er at variansen øker med middelverdien, og en slik økning indikerer at modellen ikke er en god representasjon av dataene. *Global stat* er en samlet vurdering av avvikene fra de fire premissene.

`Gvlma` består av hypotesetester for de fire premissene og den samlede vurderingen, der nullhypotesene er ikke-avvik fra premisset, altså symmetri, normal kurtose, linearitet og invariant varians (homoskedastisitet). Funksjonens standardvalg for signifikansnivå (α) er 0,05. På samme måte som for Shapiro-Wilks normalitetstest for enkeltutvalg (7.2.2.2) kan dette valget diskuteres, men jeg følger standardvalget fra Peña og Slate og det jeg oppfatter som vanlig praksis på området på samme måte som for Shapiro-Wilks normalitetstest. P-verdier over α betyr at vi vurderer det som sannsynlig at distribusjonen til de aktuelle populasjonene ikke avviker fra premisset for anova. I eksemplet over ser vi at alle premisser er oppfylt, og `gvlma` rapporterer dessuten dette med teksten "`Assumptions acceptable`", mens p-verdier under α resulterer i en tekst som signaliserer at premissene ikke er oppfylt.

De fleste modeller i kapittel 9, 10, 11 og 12 gir p-verdier over α, eventuelt på logaritme- eller logit-transformerte responsvariabler (7.2.2.5 og 7.3.2.3), men i noen tilfeller rapporterer `gvlma` p-verdier som er litt under det valgte signifikansnivået. Dette betyr at det aktuelle premisset for anova ikke er oppfylt, og at resultatet fra anova dermed ikke er gyldig. I 7.2.3.1 drøfter jeg hvordan mindre avvik fra normalitet innvirker på gyldigheten av t-testen. Anova er matematisk en utvidelse av t-testen for sammenligning av flere enn to utvalg, og den underliggende matematikken i de to analysetypene er identisk. På bakgrunn av diskusjonen om t-testens gyldighet mener jeg derfor det er trygt å regne med at også anova er relativt robust mot mindre avvik fra normalitet så lenge flere av følgende premisser er oppfylt: store utvalg, like store utvalg, moderat skjevhet og samme type skjevhet. Jeg bruker derfor i visse tilfeller denne argumentasjonen til å gjengi og bruke resultater fra anova selv om p-verdien er litt lavere enn signifikansnivået.

Det ser ut til å være vanlig prosedyre å teste bare den minimale adekvate modellen etter trinnvis modellreduksjon (Gries, 2009, s. 252-283). Dette synes å ha den ulempen at man ikke kjenner premissene for modellene som blir bygget undervegs i reduksjonsprosessen, og når kriteriet for avgjørelser i reduksjonsprosessen er knyttet til p-verdier for prediktorene i

modellen, er det fare for at man vurderer p-verdien i en modell som ikke tilfredsstiller premissene. Ideelt sett burde man derfor etter min mening teste premissene for hvert trinn i modellreduksjonen. Dette blir imidlertid en temmelig omstendelig prosess, og vurderingen av denne praksisen blir raskt en ganske stor diskusjon om forholdet mellom trinnvis modellreduksjon og trinnvis modellutvikling som ligger utenfor rammene for denne avhandlingen. Jeg har valgt å forholde meg til vanlig praksis ved å bare benytte `gvlma` på den minimale adekvate modellen, og tror ikke dette øker risikoen for modeller med svak validitet nevneverdig. (Men se også diskusjonen i 7.3.1 om risikoen for å overse adekvate modeller ved den trinnvise reduksjon av den maksimale modellen.)

I de tilfellene modellreduksjonen resulterer i nullmodellen, testes premisset for nullmodellen med en vanlig Shapiro-Wilks-test på responsvariabelen i variansanalysen, altså differanseverdien for den stilistiske variabelen eller dens log- eller logit-transformerte verdier.

Når jeg rapporterer fra variansanalysene, rapporterer jeg kort utfallet av `gvlma`-analysen, mens den fullstendige utskriften fra `gvlma` til hver analyse befinner seg i appendiks A4 og kan inspiseres der.

7.2.2.5 Logaritmetransformasjon

I denne avhandlingen brukes logaritmer til to beslektede formål, nemlig i forbindelse med statistiske analyser og som et virkemiddel i noen diagrammer.

Logaritmen av et tall *a* er det tallet *b* som logaritmebasen (for eksempel 2) må opphøyes i for å få tallet *a*.

$$\log_2(a) = b * 2^b = a$$

Vanlige logaritmebaser er 2, 10 og Eulers konstant *e*, men faktisk betyr det for formålene i denne avhandling ingenting hvilken logaritmebase man bruker, så lenge man bruker den samme basen hele tiden. Logaritmefunksjonen `log` i R bruker *e* som standardverdi, og jeg bruker derfor som oftest denne logaritmebasen i transformasjonene i avhandlingen; i illustrasjonseksemplet under bruker jeg 2, fordi det da blir enklere å følge utregningene.

Definisjonen av logaritme medfører at kvotienten mellom tallene er kongruent med differansen mellom logaritmen av tallene. Hvis kvotientene av to par av tall er de samme, er altså differansene mellom logaritmene også de samme. Dette er den sentrale egenskapen ved logaritmen som gjør den egnet til transformering av en del typer naturlig forekommende data.

(76) $a : b = c : d * \log(a) - \log(b) = \log(c) - \log(d)$
(77) $8 : 4 = 16 : 8 = 2 * \log_2(8) - \log_2(4) = \log_2(16) - \log_2(8) = 1$

En del variabler har høyreskjev distribusjon, og i en del tilfeller har dette sammenheng med at økning i variabelverdien bedre beskrives av forholdstall enn av absolutte differanser. Det

vil si at det for eksempel er mer naturlig å omtale en tekst som dobbelt så lang som en annen, enn å si at den er 400 ord lengre. Tilsvarende ville det være en mer naturlig hypotese at alle elever skriver en halv gang lengre på tastatur enn for hånd, enn at alle elever skriver 200 ord lengre. For slike variabler er distribusjonen vanligvis ikke normal, mens distribusjonen av logaritmene av verdiene gjerne er normal, ofte kalt en lognormal distribusjon. I så fall kan vi regne ut logaritmeverdier av de opprinnelige observasjonsverdiene og benytte parametriske tester på logaritmeverdiene i stedet for på de opprinnelige verdiene. I slike tilfeller sier vi at de opprinnelige tallene har gjennomgått logaritmetransformering eller log-transformering.

Også når data skal presenteres grafisk, kan det være gunstig enten å presentere logaritmeverdiene i stedet for råverdiene eller å bruke et logaritmisk aksesystem. Oftest er dette aktuelt når de statistiske testene utføres på logaritmeverdier, men noen ganger kommer distribusjonens natur bedre frem når den fremstilles logaritmisk – uavhengig av hva slags statistisk test man skal utføre.

I noen tilfeller benytter jeg logit-transformasjoner (i motsetning til logaritmetransformasjoner). Denne typen transformasjon er omtalt i 7.3.2.3 under omtalen av anova-modellering.

7.2.3 Sentraltendens (og varians)

I 7.2.1 nevnte jeg at vi ofte er interessert i å teste en "forskjell" i en variabel mellom to grupper, men jeg presiserte ikke hva en slik forskjell kan bestå i. Variabler kan nemlig være forskjellige på flere måter. En vanlig forskjell å teste på er forskjell i sentraltendens, altså et slags midtpunkt eller representativt punkt i distribusjonen, som middelverdi eller median. Middelverdi eller median trenger riktignok ikke å være en typisk verdi, men kan likevel regnes å representere en distribusjon. Men også fasongen på to distribusjoner kan være ulik, for eksempel med tanke på hvor stor spredningen er, eller hvilken veg en eventuell skjevhet går.

7.2.3.1 T-test for to uavhengige utvalg

En vanlig test for forskjell i sentraltendens er Welch' t-test (Baayen, 2008, s. 79), som er en parametrisk test som forutsetter normalfordeling. Dersom populasjonene er normalfordelte, er null-hypotesen til Welch' t-test at middelverdiene er like; det vil si at et signifikant resultat innebærer at vi konkluderer med at middelverdiene i de to populasjonene som utvalgene stammer fra, er forskjellige. Dersom populasjonene *ikke* tilfredsstiller premisset for testen, vil for det første risikoen for type-1-feil og type-2-feil endres (7.2.2.2), men dessuten vil et eventuelt signifikant resultat ha uviss årsak. En annen type distribusjonsforskjell enn forskjell i middelverdi kan ha forårsaket den lave p-verdien, og vi vet ikke hva slags forskjell eller forskjeller det kan være snakk om.

Howell (2007) peker på at det er en del diskusjon rundt spørsmålet om hvor sårbar t-testen egentlig er for brudd på premissene. Som bakgrunn for denne diskusjonen ligger sentralgrenseteoremet, som Howell (s.180) oppgir en av flere definisjonsvarianter av:

> Given a population with mean μ and variance σ^2, the sampling distribution of the mean (the distribution of sample means) will have a mean equal to μ, [and] a variance equal to σ^2/n [...]. The distribution will approach the normal distribution as n, the sample size, increases.

Sentralgrenseteoremet fastslår altså at utvalgs middelverdi vil være normalfordelt for tilstrekkelig store utvalg fra enhver populasjon. Boneau (1960) drøfter t-testens normalitetspremiss inngående og sier at (s. 60):

> it can be shown that if one samples from any two populations for which the Central Limit Theorem holds, (almost any population that a psychologist might be confronted with), no matter what the variances may be, the use of equal sample sizes insures that the resulting distribution of t's will approach normality as a limit.

Ettersom definisjonen av teoremet selv innebærer at det holder for alle distribusjoner, er det litt uklart hva Boneau egentlig mener med "almost any".

Boneau (1960) viser ved eksperimenter at visse typer av brudd på premissene har liten innvirkning på p-verdiene fra Students t-test. Eksperimentene hans er riktignok begrenset til situasjoner der middelverdien er lik i de to populasjonene; det er med andre ord kun situasjoner som kan føre til type-1-feil som er vurdert, og ikke situasjoner som kan føre til type-2-feil. Men under den forutsetningen viser han at Students t-test er temmelig robust dersom utvalgene er like store, den antatte distribusjonen har omtrent samme form, eventuell skjevhet er moderat og skjeve distribusjoner har lik varians. Ved utvalg som er av størrelse 30 eller større mener han at Students t-test er nærmest upåvirket av avvik fra premissene, mens N = 15 er tilstrekkelig dersom eventuelle avvik ikke er "ekstreme" (s. 60):

> Thus it would appear that the t test is functionally a distribution-free test, providing the sample sizes are sufficiently large (say, 30, for extreme violations) and equal.

Han peker dessuten på Welch' tilpasning av Students t-test og dens toleranse for avvik fra premisset om lik varians. Nyere lærebøker som Levshina (2015, s. 88) og Urdan (2010, s. 53) støtter seg antagelig på Boneau og tilsvarende eksperimenter når de uten å oppgi kilder eller matematiske argumenter hevder at nettopp N = 30 er tilstrekkelig utvalgsstørrelse for å se bort fra premisset om normalitet i t-testen. Ettersom de aktuelle utvalgene i analysene i denne avhandlingen stort sett er like store (se 7.3.2.2), som regel er minst N = 30 og alltid minst N = 15, støtter jeg meg også på Boneau når jeg rapporterer enkelte resultater fra Welch' t-test selv om utvalgene er moderat skjeve og Shapiro-Wilks normalitetstest rapporterer p-verdier som er noe lavere enn 0,05. Jeg følger dessuten den samme praksisen for anova (7.3.2.1), ettersom de underliggende prinsippene og matematiske beregningene i anova er de samme som i Students t-test.

En variant av t-testen tester på parede data, altså der de to delutvalgene har verdier for de samme individene slik at hvert individ er representert én gang i hvert utvalg, og man vet hvilke to observasjoner som hører til samme individ. For eksempel kan man i aksjonsforskning teste de samme elevene på den samme egenskapen før og etter et tiltak; en paret test vil vurdere endringer i hvert individ og ikke bli påvirket av den interindividuelle variasjonen i utvalget, som gjerne kan være stor. Dette øker testens styrke til å avdekke

reelle tendenser i populasjonen. I denne avhandlingen er det skriveverktøyet som representerer tiltaket i den parede designen, og en paret t-test utnytter designen til å evaluere forskjeller i individene mellom de to skriveverktøyene uten å bli forstyrret av den interindividuelle variasjonen. I praksis viser det seg i eksperimentet i denne avhandlingen at den intraindividuelle variasjonen er så stor at den parede t-testen ofte har liten eller ingen styrkefordel over den uparede.

Den parede versjonen av t-testen fungerer på den måten at distribusjonen av par-differansene blir sammenlignet med en uniform nulldistribusjon, altså et konstruert utvalg som består av bare null-verdier. Det innebærer at premisset om normalfordeling gjelder distribusjonen av differanseverdier. Dessuten er det en forutsetningen at variansen ikke varierer med variabelverdiene. Dette er et premiss som vil være brutt der variablene har preg av å være forholdstall, og i disse tilfellene må variablene transformeres før testing eller utregning av differanseverdier (7.2.2.5 og 7.3.2.3).

I denne avhandlingen bruker jeg to versjoner av t-testen, nemlig paret t-test og Welch' t-test. I det følgende bruker jeg alltid betegnelsen "t-test" om en av disse, ettersom jeg aldri benytter Students t-test.

Jeg oppgir effektstørrelser som Cohens *d*, som er forskjellen i middelverdier målt i gjennomsnittet av de to utvalgenes standardavvik. Se appendiks E1.

7.2.3.2 Kolmogorov-Smirnov-test

Kolmogorov-Smirnov-testen (KS-test) er en ikke-parametrisk test som tester generelt for *ulikhet i distribusjon* (Baayen, 2008, s. 78-79; Crawley, 2007, s. 317; Gries, 2009, s. 159-165) ved å sammenligne akkumulerte verdisummer av rangerte data. KS-testen nevnes ofte som et ikke-parametrisk alternativ til t-testen, men testens nullhypotese er ikke lik sentraltendens men lik distribusjon. Ulike utvalg kan altså gi positivt resultat med denne testen selv om både middelverdi og varians er lik, dersom fordelingene for eksempel er skjeve i hver sin retning. Et positivt resultat av testen sier dermed ingenting om hvorvidt forskjellen skriver seg fra forskjell i sentraltendens, varians, skjevhet eller kurtose.

Formålet med å bruke en KS-test kan være som ikke-parametrisk test der premissene for parametriske tester ikke er oppfylt, eller der konsekvensene av brudd på premissene er usikre. Formålet kan imidlertid også være å avdekke om andre forskjeller i distribusjonen enn sentralverdi og varians er signifikante. Siden KS-testen er fordelingsfri, kan den også brukes som en rask metode for å avsløre om forskjeller kan være signifikante, uten å måtte gjøre vurderinger av normalitet eller varians først. KS-testen har ofte heller ikke så mye dårligere styrke enn t-testen, ettersom mangelen på parametriske effekter kompenseres med følsomhet for andre egenskaper ved distribusjonen, og ettersom KS-testen benytter variablenes verdier og ikke bare deres rangering, slik Wilcoxons rangsum-test gjør (se 7.2.3.3).

7.2.3.3 Wilcoxons rangsum-test

Wilcoxons rangsum-test, som også blir kalt Mann-Whitneys U-test, er en ikke-parametrisk test som i likhet med Kolmogorov-Smirnov-testen tester for ulikhet i distribusjonen, men som er særlig følsom for ulikhet i sentraltendens. Wilcoxon-testen har gjerne noe lavere styrke enn t-testen (Baayen, 2008, s. 77)og kan dermed sees som en konservativ test som senker risikoen for type-1-feil. Ettersom kun rangeringsdata blir brukt i beregningene, er ikke Wilcoxon-testen sårbar for utliggere, i motsetning til KS-testen og særlig t-testen.

I likhet med t-testen finnes en paret variant også av Wilcoxons rangsum-test. Den parede varianten tester om *differansene* mellom to datasett tenderer til å være større eller mindre enn null.

Selv om Wilcoxon-testen gjerne blir omtalt som et ikke-parametrisk alternativ til t-testen, er det viktig å være klar over at de ulike beregningsmåtene som benyttes i testen, innebærer at de ikke har akkurat samme nullhypotese og dermed ikke tester akkurat samme hypotese. De to testene kan derfor gi ulikt resultat på forskjellige distribusjoner, og ikke nødvendigvis bare på den måten at t-testen alltid gir lavere p-verdier. I visse typer av parede tester kan kanskje en rangeringsbasert test gi et bilde av situasjonen som har tettere sammenheng med det forskningsspørsmålet vi faktisk er interessert i. Samtidig vil jeg nevne at det er en viss fare forbundet med å bruke rangeringsbaserte metoder på frekvensverdier som skriver seg fra fenomener med få forekomster i hver tekst; interaksjon mellom frekvensverdiene og verdiene for det som er brukt som målestokk, for eksempel tekstlengde, kan føre til at små forskjeller i frekvensverdier kan gi kunstig sterk effekt på resultatet, og man risikerer å rapportere et positivt "funn" som i realiteten representerer svaret på et annet forskningsspørsmål enn det man hadde som intensjon. For slike variabler kan det tenkes at en t-test gir mer valide resultater, selv om populasjonene ikke oppfyller premisset om normalitet. (Se også diskusjonen i 7.2.3.1.)

7.2.4 Korrelasjon

Jeg bruker to ulike tester for korrelasjon, en parametrisk og en ikke-parametrisk. Pearsons produkt-momentkorrelasjon med korrelasjonskoeffisient *R* og Spearmans rangeringsbaserte korrelasjon med korrelasjonskoeffisient rho (eller *R*s). (Det er vanlig å bruke den greske bokstaven ρ som betegnelse for Spearmans korrelasjonskoeffisient, men jeg skriver bokstavnavnet med latinske bokstaver for å unngå feillesninger av ρ som p.) Pearsons korrelasjonstest bygger på de samme forutsetningene som t-testen, altså særlig normalfordelt populasjon, mens Spearmans korrelasjonstest er en fordelingsuavhengig (Baayen, 2008, s. 91), ikke-parametrisk test (Dodge, 2010, s. 502-505) som gjør sine beregninger på grunnlag av rangeringsverdier og kan brukes på ikke-normalfordelte variabler eller for å nøytralisere utliggere. Mange av de vurderingene av parametriske kontra rangeringsbaserte tester som jeg er inne på i omtalene av henholdsvis t-test og Wilcoxon-test, gjelder også for Pearsons og Spearmans korrelasjonstester.

På samme måte som andre hypotesetester resulterer korrelasjonstestene i p-verdier som kan sammenlignes med α. Korrelasjonskoeffisientene er derimot effektmål som betegner styrken i korrelasjonen uavhengig av utvalgets størrelse. I en del sammenhenger i avhandlingen oppgir jeg korrelasjonskoeffisienter som mål på sammenheng mellom variabler uten å bruke beregningen som grunnlag for en signifikanstest. Noen ganger beregner jeg også korrelasjonskoeffisienter for samtlige 120 tekster i korpuset, selv om 2 og 2 tekster er skrevet av samme elev og disse dermed ikke utgjør bare uavhengige observasjoner. Så lenge disse beregningene ikke benyttes til konklusjoner om signifikans, kan slike korrelasjonskoeffisienter likevel brukes til å gi et godt og ganske valid inntrykk av sammenhenger mellom variabler.

7.2.5 Signifikansnivå

Jeg diskuterer i 7.2.1 valg av signifikansnivå og slår fast at jeg bruker $\alpha = 0{,}05$ i denne avhandlingen.

Relatert til valg av signifikansnivå er valg av enhalet kontra tohalet testing. Jeg har utført alle signifikanstester i denne avhandlingen som tohalede tester. Formelt innebærer dette at hypotesen for for eksempel en t-test er at det finnes en forskjell mellom to (eller flere) utvalg, og ikke at et spesifisert utvalg har høyere verdier enn det andre. Bakgrunnen for dette valget er kompleksiteten i den overordnede hypotesen (5.1). Siden jeg antar at mange ulike faktorer påvirker variablene på ulike måter, har jeg heller ingen prediksjon om en konkret retning på effekten av prediktorvariabelen skriveverktøy. Dette er et konservativt valg som nok kan kritiseres for å øke risikoen for type-2-feil, altså å bekrefte falske null-hypoteser, ettersom p-verdien for en tohalet test er dobbelt så høy som for en enhalet test (i symmetriske distribusjoner). Jeg kompenserer noe for denne økte risikoen ved å oppgi p-verdiene for ikke-signifikante resultater, slik at leseren selv kan vurdere om en enhalet test ville ha gitt et annet resultat. Imidlertid vil en slik form for implisitt kompensering for tohalet testing ikke være mulig i den trinnvise reduksjonen av anova-modeller, ettersom nesten-signifikante faktorer der kan bli fjernet når p-verdien blir brukt som kriterium, slik jeg gjør.

På bakgrunn av argumentasjonen over kan denne studien sies å ha et eksplorativt preg. Hypotesen predikerer ikke konkrete effekter, og både det teoretiske og det empiriske grunnlaget for hypotesene er relativt svakt, ettersom mesteparten av både teori og empiri ikke er knyttet direkte til skriveverktøy. Det at studien er delvis eksplorativ, er også et argument for å bruke p-verdiene fra de enkelte testene og $\alpha = 0{,}05$ uten å kompensere for det som er kjent som *familywise error rate* (FWER), altså den økte risikoen for type-1-feil som oppstår når man utfører multiple tester på det samme materialet. Totalt 13 anova-analyser utgjør det som jeg anser som hovedanalysene i studien, og korrigering for FWER ville dermed resultert i en betraktelig økning i risiko for type-2-feil, noe som ville være uheldig i et eksplorativt perspektiv. I etterkant av de 13 anova-analysene forsøker jeg imidlertid å samle de 13 variablene i en helhetlig, multivariat analyse, nettopp med tanke på justere den eventuelle overrapporteringen av positive funn som de gjentatte testene av materialet gir.

7.3 Statistiske modeller

Metodene som er omtalt i 7.2.3 og 7.2.4, omfatter bare metoder for hypotesetesting. I dette avsnittet omtales metoder for statistisk modellering. De fleste analysene i denne avhandlingen bruker metoder av denne typen. Metodene som er omtalt, er grunnleggende og allment kjent, og dette avsnittet gir ingen generell innføring i statistisk modellering; omtalene baserer seg hovedsakelig på (Baayen, 2008; Crawley, 2005, 2007; Gries, 2009), og mer utfyllende forklaringer kan finnes her eller for eksempel i (Aitkin, Francis, Hinde, & Darnell, 2009; Faraway, 2005; Field, et al., 2012; Howell, 2007).

I en modell kalles variablene som *i modellen* sees som de avhengige variablene, 'responsvariabler', mens variablene som sees som uavhengige variabler og kan *predikere* responsen, kalles 'prediktorer'. Det er viktig å merke seg at det ikke nødvendigvis er slik at det er et kausalt forhold fra prediktorer til respons i den virkelige verden; det kan være nyttig å modellere variabler på denne måten i en analyse selv om de ikke står i noe direkte kausalt forhold, eller selv om kausaliteten går i den motsatte retningen. I denne avhandlingen er de fleste modellene univariate; det vil si at de har bare én responsvariabel. De kan imidlertid gjerne ha flere prediktorer, og i de fleste tilfellene har de det. Hypotesetestene jeg har omtalt så langt, kan sees som forenklede versjoner av de statistiske modellene jeg omtaler i dette delkapitlet, og betegnelsene 'responsvariabel' og 'prediktor' eller 'prediktorvariabel' kan derfor også brukes i sammenheng med hypotesetestene.

Jeg begynner med å omtale ordinær lineær regresjon med kontinuerlig responsvariabel og kontinuerlige prediktorer. Mange av begrepene og metodene som senere blir benyttet i omtalen av variansanalyse, blir introdusert og forklart her.

7.3.1 Lineær regresjon

I en lineær regresjonsmodell søker man å forklare variasjonen i en kontinuerlig responsvariabel (avhengig variabel) med variasjonen i én eller flere kontinuerlige prediktorer (uavhengig variabel). Dersom det er flere enn én prediktor, kan man velge å betrakte også interaksjonen mellom prediktorer som prediktorer i modellen. Det er to hovedprinsipper for regresjonsanalyse, der den ene er basert på at man tar utgangspunkt i en maksimal modell og gjennomfører en prosedyre for å redusere denne modellen til den minste modellen som har det man vurderer som tilstrekkelig forklaringskraft, og den andre er basert på at man tar utgangspunkt i en minimal modell og gjennomfører en prosedyre for gradvis å bygge opp en mer omfattende modell som er den minste modellen som har det man vurderer som tilstrekkelig forklaringskraft. De to prosedyrene har ulike fordeler og ulemper, og de kan også kombineres. Jeg har valgt prosedyren som benytter trinnvis reduksjon, ettersom risikoen for at prosedyren skal overse relevante modeller synes å være minst for denne fremgangsmåten. Den synes dessuten å være den metoden som er mest utbredt, trolig av samme grunn.

Framgangsmåten for trinnvis reduksjon er at man konstruerer en maksimal modell som regel med flere prediktorledd. Prediktorledd kan være prediktorer, interaksjoner mellom

prediktorer og eventuelt også ikke-lineære ledd som for eksempel kvadratledd, og modellen tilegner hvert av leddene en p-verdi som gjenspeiler hvorvidt det aktuelle prediktorleddet bidrar til en positiv tendens i modellen. Man reduserer deretter den maksimale modellen ved å fjerne prediktorledd som i liten grad bidrar til å forklare variasjon i responsvariabelen. Det er flere alternative kriterier for fjerning av prediktorledd; vurdering av modellalternativenes *Akaike Information Criterion* (AIC) er en anerkjent metode, men jeg følger Gries (2009, s. 260) og bruker sammenligning av leddenes p-verdier som kriterium. Prediktorledd fjernes da ett ad gangen etter følgende prinsipper:

1. Bare ledd med $p > \alpha$ fjernes.
2. Et interaksjonsledd fjernes før leddene som inngår i interaksjonen.
3. Et ledd med flere interaksjoner fjernes før ledd med færre interaksjoner dersom samme prediktor inngår i begge leddene.
4. Blant leddene som tilfredsstiller de tre kriteriene over, fjernes leddet med høyest p-verdi først.

Ledd fjernes iterativt inntil det ikke finnes flere ledd med $p > \alpha$ som kan fjernes. Den modellen som da gjenstår, kalles den minimale adekvate modellen. Denne modellen har (som alle de foregående modellene i den trinnvise reduksjonsprosessen) en p-verdi som gjenspeiler hvorvidt det finnes tendenser i modellen som avviker fra nullhypotesen, men uten å peke på hvilke delutvalg det eventuelt finnes forskjeller mellom. Man kan da ha tre situasjoner:

a) En modell uten gjenværende prediktorer. Dette kalles nullmodellen.
b) En modell med $p < \alpha$ med minst ett gjenværende ledd med $p < \alpha$.
c) En modell med $p > \alpha$, men der ingen av leddene kan fjernes etter reglene over.

Ved situasjon c) har man et ikke-signifikant resultat av analysen. Ved situasjon a) er den videre fremgangsmåten avhengig av forskningsspørsmålet; se 7.3.2.

Ettersom interaksjoner mellom tre prediktorer deler utvalget på 60 differanseverdier inn i 8 delutvalg av gjennomsnittlig størrelse $N = 7{,}5$, tar jeg i anova-analysene utgangspunkt i maksimale modeller som er begrenset til interaksjoner mellom to prediktorer. Med 4 prediktorer innebærer dette totalt 6 interaksjoner. Interaksjoner med tre prediktorer kan også være svært vanskelige å tolke, og dette kan være et argument i seg selv for å avgrense nivået av interaksjoner til 2.

7.3.2 Multifaktoriell variansanalyse

Den mest brukte analysemetoden i denne studien er multifaktoriell variansanalyse, gjerne kalt anova. Prinsippet for modellbygging og trinnvis reduksjon er det samme som for regresjonsanalyse.

7.3.2.1 Vanlig anova med uavhengige observasjoner

Variansanalyse har likhetstrekk med regresjonsanalyse, men i variansanalysen er prediktorene nominale. I alle anova-analysene i denne avhandlingen er prediktorene dessuten dikotome.

Siden problemstillingen i denne avhandlingen er knyttet til hvorvidt variabler har ulik verdi i to ulike skrivesituasjoner, er responsvariablene i denne avhandlingen i de fleste tilfellene *forskjell* i en variabelverdi mellom tasteteksten og håndteksten. Forskjellen kan være differanser mellom frekvensverdier, differanser mellom logaritmetransformerte frekvensverdier (7.2.2.5) eller differanser mellom logit-transformerte verdier (7.3.2.3). Det vil altså si at N i modellen er antall elever og ikke antall tekster. Differansen blir alltid regnet ut med tasteverdien som minuend og håndverdien som subtrahend, altså som vist i formelen i (78). Det innebærer at variabelverdiene blir negative for elever med høyere verdier i håndteksten enn i tasteteksten.

(78) $var1_{diff} = var1_{tast} - var1_{hånd}$

Analysen tar utgangspunkt i en maksimal modell med den aktuelle differanseverdien av en leksikosyntaktisk variabel som responsvariabel og 4 elevvariabler og deres interaksjoner begrenset til 2 nivåer som prediktorvariabler. Den ordinære maksimale modellen i denne avhandlingen bruker de fire dikotome prediktorene kjønn, skriveferdighet, total tekstlengde større eller mindre enn medianen, og kvotienten av tekstlengdene større eller mindre enn medianen. (Se forklaring av prediktorene i 7.3.2.2 nedenfor.) Et eksempel på en maksimal modell er (79), som modellerer differansen i ordvariasjonsindeksen OVIX (10.3.2):

(79) `lm(lexD$ovix ~ (kjønn+ferdighet+lengde+forskjell)^2)`

I alle variansanalysene i de følgende kapitlene blir de fire prediktorene gitt navn som vist i formelen i (79). Denne modellen blir så trinnvist forenklet gjennom å fjerne det leddet som bidrar minst til å forklare variasjonen i responsvariabelen, med samme fremgangsmåte som blir forklart for lineær regresjon i 7.3.1.

Dersom den minimale adekvate modellen er signifikant, kan man inspisere de enkelte leddenes p-verdier for å avdekke hvilke av dem som har signifikant innvirkning på responsvariabelen. Ofte omfatter dette alle de gjenværende leddene, men det er også mulig at ikke-signifikante ledd gjenstår i modellen fordi de inngår i en signifikant interaksjon.

Dersom den minimale adekvate modellen inneholder interaksjoner, kreves det ytterligere analyse for å avdekke egenskapene til den signifikante interaksjonen. En vanlig post-hoc-test for anova med interaksjoner av denne typen, er Tukeys HSD-test (for *honestly significant differences*). Denne testen peker ut hvilke delutvalg det finnes signifikante forskjeller mellom, og den tar i beregningene hensyn til FWER (7.2.5) og korrigerer for dette. I enkelte tilfeller finner Tukeys HSD-test *ingen* signifikante forskjeller selv om interaksjonen i seg selv er signifikant. Dette kan det være flere årsaker til, men en viktig årsak er den interaksjonseffekten som vi ser et eksempel på i analysen av korte subklaususer i 11.1.3.5.

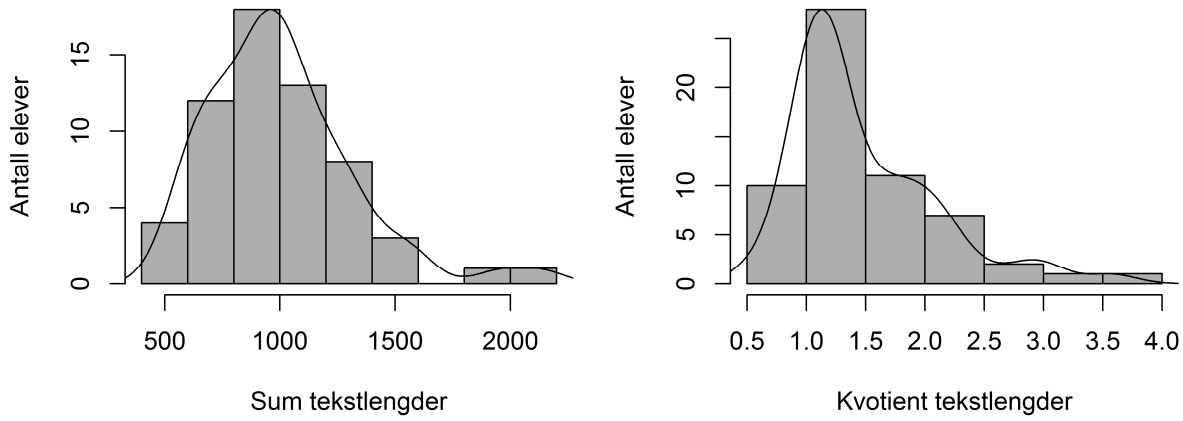
Tukeys HSD-test forutsetter at delutvalgene er omtrent like store (Baayen, 2008, s. 106), noe de er i denne studien.

Nullmodellen kan også inneholde et signifikant resultat, som kan gå fram av den dokumentasjonen som R presenterer av modellen, eller kan testes spesielt med en paret t-test på responsvariabelen. En signifikant nullmodell innebærer i denne studien at det er en allmenn effekt av skriveverktøyet på den aktuelle leksikosyntaktiske variabelen.

7.3.2.2 Prediktorer og dikotomisering av kontinuerlige parametre

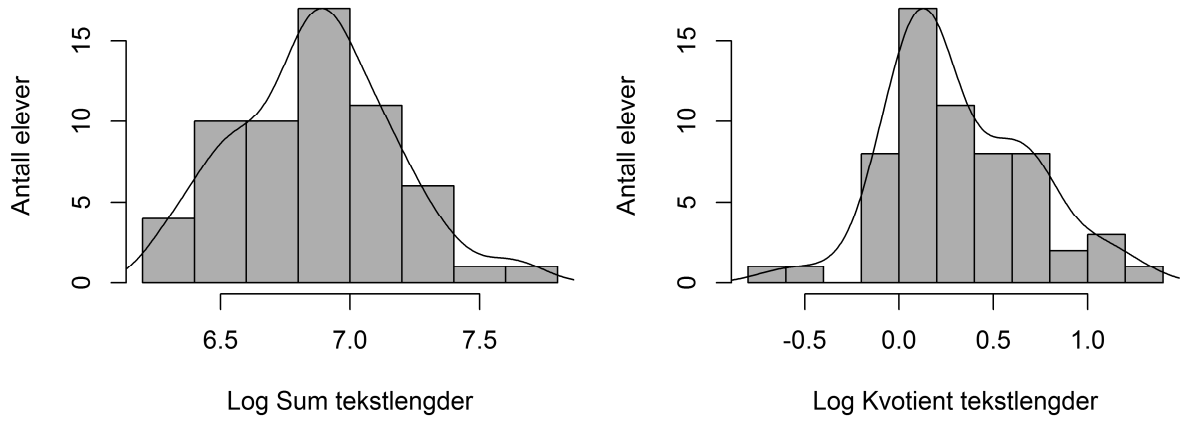
De 4 prediktorene som blir mest benyttet i analysene i denne avhandlingen, er kjønn, skriveferdighet, total tekstlengde og forskjell i tekstlengde. Kjønn er i utgangspunktet dikotom. Skriveferdighet er målt som elevenes norskkarakter (5.3.2), men siden nesten alle elevene har 4 eller 5 i norskkarakter, er det naturlig å dikotomisere også denne variabelen, og jeg gjør det ved å regne elever med 5 og 6 i karakter som "sterke" elever og elever med 3 eller 4 i karakter som "middels" elever. De to siste prediktorene er imidlertid inherent kontinuerlige. Dette gjelder summen av tekstlengde i elevens to tekster, og det gjelder forholdstallet (kvotienten) mellom lengdene på elevens to tekster. Total tekstlengde regnes i antall ord, altså heltall, og er dermed prinsipielt diskret, men for alle praktiske formål kan vi se på den som kontinuerlig. Forholdstallene er rasjonelle og både i praksis og prinsipielt kontinuerlige.

Prediktorene fremstår altså som en kombinasjon av 2 dikotome og 2 kontinuerlige parametre, og slike kombinasjoner krever kovariansanalyse (ancova) (se f.eks. Crawley (2007, s. 489-)). De to kontinuerlige parametrene er imidlertid litt problematiske. Ingen av dem er normalfordelt, og selv om begges logaritmiske transformasjoner rapporteres som normalfordelt av Shapiro-Wilks normalitetstest, fremstår ingen av dem heller som typisk lognormale.



Figur 7-2: Histogrammer med tetthetskurver for henholdsvis summen av tekstlengder (til venstre) og for forholdstallet mellom tekstlengder (til høyre). Shapiro-Wilks normalitetstest gir lave p-verdier ($W \approx 0{,}92$, $p < 0{,}001$; $W \approx 0{,}84$, $p < 0{,}001$).

Figur 7-2 viser at for total tekstlengde er det særlig to elever med høye verdier som påvirker normaliteten. Forskjellen i tekstlengde virker i utgangspunktet mer typisk lognormal, men figur 7-3 viser at p-verdien fra Shapiro-Wilks normalitetstest for de logaritmetransformerte verdiene fortsatt er temmelig lav, og tetthetskurven er ikke typisk normal. Transformeringen resulterer faktisk også i to nye utliggere til venstre. Logaritmetransformering av totallengdene resulterer imidlertid i en ganske normal fordeling.



Figur 7-3: Histogrammer med tetthetskurver for henholdsvis logaritmetransformerte totallengder (til venstre) og lengdeforskjeller (til høyre). Shapiro-Wilks normalitetstest gir $W \approx 0{,}986$, $p \approx 0{,}71$ og $W \approx 0{,}965$, $p \approx 0{,}08$.

At parameteren tekstlengdeforskjell hverken er typisk normal eller typisk lognormal, medfører problemer om den skulle inkluderes i en kovariansanalyse, og det gjør det også utfordrende å fortolke eventuelle positive resultater. Det er selvfølgelig mulig at det finnes andre transformasjonsmåter som vil transformere variabelen til normalfordeling, men jeg har ikke funnet noen. Dessuten bør en eventuell transformasjon gjenspeile den konseptuelle distribusjonen, og ikke bare tilfeldigvis ende opp i et utvalg som Shapiro-Wilk-testen vurderer som normalfordelt.

For å forenkle både analyse og fortolkning har jeg derfor valgt å dikotomisere tekstlengdeforskjell rundt medianverdien, og jeg har valgt å behandle totallengde på samme måte. Det vil altså si at jeg har gjort dem om til dikotome variabler som hver deler elevene inn i to like store grupper. Den ene variabelen deler inn elevene i en gruppe som skriver *kort*, og en gruppe som skriver *lang*t, altså kortere eller lengre enn medianpunktet i distribusjonen. Den andre variabelen deler inn elevene i en gruppe som har *liten* forskjell mellom lengden i de to tekstene, og en gruppe som har *stor* forskjell. Det vil si at de skriver mye lengre tekster på tastatur enn for hånd (se 8.4.2). Som dikotome variabler kan de inngå i en ordinær anova-analyse med bare binære prediktorer i stedet for en mer kompleks ancova-analyse.

Det er en viss interaksjon mellom de to tekstlengdevariablene, noe som fører til at den balansen som gjelder for kjønn og ferdighet, ikke gjelder fullt ut for de to tekstlengdevariablene:

		Tekstlengdeforskjell		
		Liten	Stor	**Sum**
Total tekstlengde	Kort	18	12	30
	Lang	12	18	30
	Sum	**30**	**30**	**60**

Det er også noen skjevheter knyttet til kjønn og ferdighet:

Kjønn og Ferdighet	Total tekstlengde	Tekstlengdeforskjell		
		Liten	Stor	**Sum**
Gutter	Kort	9	7	16
	Lang	6	8	14
	Sum	**15**	**15**	**30**
Jenter	Kort	9	5	14
	Lang	6	10	16
	Sum	**15**	**15**	**30**
Middels	Kort	11	8	19
	Lang	4	7	11
	Sum	**15**	**15**	**30**
Sterke	Kort	7	4	11
	Lang	8	11	19
	Sum	**15**	**15**	**30**

Dette innebærer altså at alle utvalg ikke er balanserte, noe som har negativ innvirkning på analysenes robusthet mot avvik fra normalitet (7.2.3.1), og som kan spille en rolle i fortolkning av resultatene av variansanalysene.

7.3.2.3 Logit-transformering

En del av responsvariablene i denne undersøkelsen er forholdstall. Et typisk eksempel er andel t-enheter med nøyaktig ett ord i forfelt. Det er flere argumenter for å ikke bruke lineær regresjon eller variansanalyse når responsvariabelen er forholdstall (se f.eks. Crawley (2007, s. 569-)):

- Responsvariabelen har begrenset utstrekning mellom 0 og 1, mens vanlig lineær regresjons- eller variansanalyse forutsetter responsvariabler uten teoretiske avgrensninger.
- Variansen er ikke konstant i verdiområdet, men lavere nær utkantene.
- Residualene er ikke normalfordelt.

Alle disse problemene kan løses ved å bruke logistisk regresjon, for eksempel med R-funksjonen `glm` (for *generalised linear models*).

Eksperimentet i denne avhandlingen er designet som et *repetert forsøk*, der den samme typen observasjon er utført to ganger på hver informant, og det primære formålet med eksperimentet er å avdekke *endring* eller forskjell mellom de to observasjonene og ikke observasjonsverdiene i seg selv. Selv om det er mulig å modellere repeterte forsøk også i logistisk regresjon, er dette en analyseform jeg ikke har satt meg inn i, og jeg har valgt en

enklere tilnærming som likevel tar hensyn til momentene til Crawley ovenfor. Jeg forklarer denne tilnærmingen i det følgende.

Ettersom responsvariabler som er sannsynlighetsverdier eller forholdstall med verdier mellom 0 og 1, ikke har konstant varians, men har asymptotiske egenskaper med hensyn til grenseverdiene 0 og 1, er det fortolkningsmessig meningsløst å beregne *forskjeller* mellom dem ved å regne ut differanser eller kvotienter. For eksempel er en differanse på 0,1 mellom 0,55 og 0,45 en vesentlig mindre betydningsfull forskjell enn en differanse på 0,1 mellom 0,11 og 0,01. Tilsvarende er en kvotient på 2 mellom 0,1 og 0,05 ikke av like stor betydning som en kvotient på 2 mellom 0,9 og 0,45. Verdienes avstand til grenseverdiene er av betydning, og denne betydningen kommer ikke fram ved differanser eller kvotienter.

I variabler som reelt er forholdstall, kan man anslå sannsynligheten p for positivt utfall i en binær observasjon, for eksempel sannsynligheten for at en t-enhet har nøyaktig ett ord i forfelt. Hvis en tekst inneholder 12 t-enheter og 3 av dem har ett ord i forfeltet, kan man anslå at sannsynligheten er 3 / 12 = 0,25 for at en t-enhet inntreffer med ettordsforfelt for den aktuelle eleven og med de aktuelle situasjonsparametrene. Til grunn for logistisk regresjon ligger såkalt logit-transformasjon (Baayen, 2008, s. 196; Crawley, 2007, s. 571-573), som er logaritmisk transformasjon av *odds* i stedet for av p-verdier. Oddsen for et utfall med sannsynlighet p er $p/(1 - p)$. I en tekst der 3 av 12 t-enheter har nøyaktig ett ord i forfelt, finnes *logit*-verdien av variabelen slik:

(80) $$logit(p) = ln\frac{p}{(1-p)} = ln\frac{0{,}25}{0{,}75} = ln\frac{1}{3} = -\,ln\,3 \approx -\,1{,}099$$

Logit-transformerte forholdstall har lineære egenskaper, i motsetning til de asymptotiske egenskapene til sannsynlighetsverdiene, og det er derfor meningsfylt å måle forskjell mellom logit-verdier ved å regne ut differansen. Vi kan dermed bruke differanse mellom logit-transformerte forholdstall som responsvariabel i et repetert forsøk med for eksempel andel t-enheter med ett ord i forfelt. Det er dette prinsippet som ligger til grunn for logistisk regresjonsanalyse med for eksempel funksjonen `glm` i R.

Det er imidlertid to problemer med formelen slik den er gjengitt over. Jeg skal demonstrere problemene og mulige løsninger gjennom et konkret eksempel, nemlig andelen av nominale subklaususer som ikke har subjunksjon. Det er vanlig å regne med at nominale subklaususer uten subjunksjon står i kontrast til nominale subklaususer med subjunksjonen \at\, ettersom man normalt kan sette inn \at\ i subjunksjonsløse nominalklaususer uten å endre klaususens mening, mens andre nominale subjunksjoner som \om\ eller \hva\ ikke kan settes inn på samme måte. En naturlig målestokk for antall nominale subklaususer uten subjunksjon er dermed antall nominale subklaususer som enten har \at\ eller mangler subjunksjon. Strengt tatt er ikke dette konseptuelt helt presist, for \at\ kan ikke strykes i alle nominale \at\-klaususer; for eksempel kan ikke \at\ strykes i frontale \at\-klaususer, og heller ikke i klaususer som står som utfylling til preposisjon eller i visse typer utbrytninger. Som illustrerende eksempel er imidlertid dette presist nok.

Det første problemet er knyttet til at logit-verdien i eksemplet i (80) er udefinert for $p = 0$ og for $p = 1$. For $p = 0$ gir formelen som resultat logaritmen av 0, som er minus uendelig. For $p = 1$ blir divisor i formelen $(1 - p) = 0$, og resultatet er udefinert. (Eventuelt kan man si at resultatet blir logaritmen av uendelig, som er uendelig, men problemet forblir det samme.)

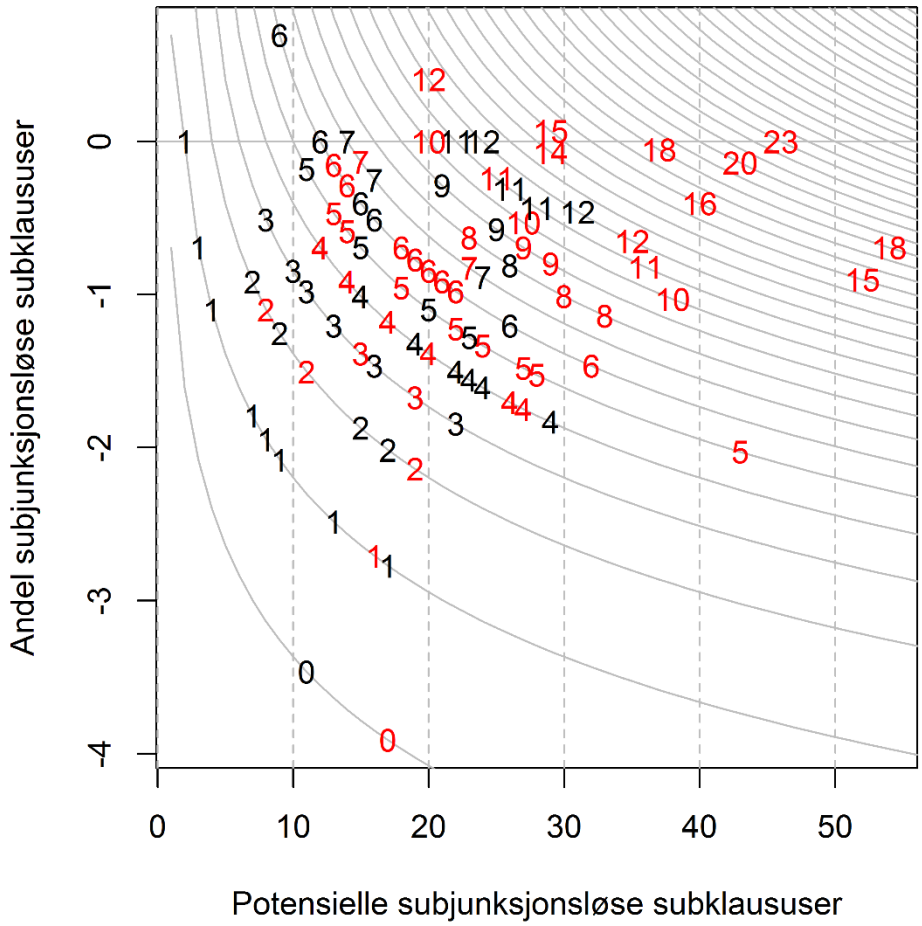
P-verdiene i formelen vil normalt regnes ut ved at antall enkeltforekomster av variabelen divideres med antall potensielle forekomster. I dette eksemplet blir *p*-verdien regnet ut etter følgende formel:

(81) $$p = \frac{kl_{uten}}{kl_{at} + kl_{uten}}$$

Resultatet av formelen over blir $p = 0$ for $kl_{uten} = 0$, altså ingen subjunksjonsløse subklaususer, og $p = 1$ for $kl_{at} = 0$, altså bare subjunksjonsløse subklaususer.

Det er på sett og vis intuitivt riktig at (80) ikke gir noe resultat for *p* lik 0 eller 1, for sannsynligheten for \at\-elidering er aldri nøyaktig 0 eller 1 i en gitt skrivesituasjon. Ettersom \at\-elidering er en del av grammatikken i moderne norsk, må vi også regne med at moderne språkbrukere vil kunne produsere tilfeller av \at\-elidering, og ettersom \at\-elidering ikke er obligatorisk, må vi også regne med at moderne språkbrukere vil kunne produsere subklaususer der \at\ ikke er elidert. Fravær av henholdsvis subjunksjonsløse nominalklaususer eller nominalklaususer med subjunksjon kan dermed sees på som et resultat av at den aktuelle teksten er for kort til at fenomenet tilfeldigvis har forekommet innenfor akkurat disse rammene, men ikke som et resultat av at sannsynligheten for utfallet er 0. Det er dermed i tekster med 0 positive forekomster rimelig å anta en *p*-verdi som representerer et sted mellom 0 og 1 positive forekomster i den aktuelle tekstlengden. Tilsvarende er det i tekster med 0 negative forekomster rimelig å anta en *p*-verdi som representerer et sted mellom 0 og 1 negative forekomster i den aktuelle tekstlengden.

Jeg har valgt å basere *p*-verdien på ⅓ forekomst i tekster uten forekomster. Tallet ⅓ er ikke teoretisk basert, men valgt slik at den logaritmiske avstanden mellom ⅓ og 1 blir større enn den logaritmiske avstanden mellom 1 og 2, som igjen er større enn avstanden mellom 2 og 3, etc. Figur 7-4 viser subjunksjonselidering for nominalklaususer og relativklaususer samlet, og figuren demonstrerer hvordan avstanden mellom kurvene øker med synkende antall potensielle omgivelser. Figuren illustrerer også hvordan transformasjonen gir høyere verdier ved færre potensielle omgivelser, også dersom antall forekomster er 0. Kurvene demonstrerer at tekster uten forekomster i gitte tilfeller kan få høyere transformerte verdier enn lengre tekster *med* forekomster, selv om eksemplet ikke viser slike konkrete tilfeller. Dette er en naturlig følge av at sannsynligheten for forekomster i en gitt skrivesituasjon ikke er 0 selv om antall forekomster i den konkrete teksten er 0.



Figur 7-4: Logit-transformerte andel subjunksjonsløse subklaususer, altså både nominalklaususer og relativklaususer. Tallene angir antall subjunksjonsløse klaususer i teksten. De grå kurvene viser de teoretiske logit-verdiene for ulike kombinasjoner av faktiske forekomster og potensielle forekomster.

Det andre problemet knyttet til logit-transformasjoner er relatert til de tilfellene der antall potensielle omgivelser er 0. I eksemplet over ville det innebære at både antall nominale subklaususer med \at\ og antall nominale subklaususer uten subjunksjon er 0. Siden nominale subklaususer er en temmelig frekvent konstruksjon i norsk, vil dette sjelden forekomme i vanlige tekster av noen lengde, men i dette tekstutvalget er det én tekst uten nominale subklaususer av disse typene, mens det laveste antallet av potensielt subjunksjonsløse subklaususer er 2, representert av 1-tallet lengst til venstre i figur 7-4. For andre typer variabler kan dette være et mer frekvent problem. Det er tre måter å håndtere slike variabler på.

- For det første kan man unnlate å analysere slike variabler kvantitativt. Dersom antall tekster med null forekomster av potensielle omgivelser er en vesentlig andel av tekstmaterialet, kan dette være den beste løsningen. Dersom mange tekster har et lavt antall potensielle forekomster, kan det også indikere at det vil være vanskelig å gi en relevant og valid analyse av variabelen. (Se også 7.2.3.3.)

- For det andre kan man utføre analysen bare på det utvalget av tekstsamlingen som har forekomster av potensielle omgivelser. Man bør i så fall vurdere hva mangelen på potensielle omgivelser skyldes, og eventuelt gjøre en frafallsanalyse.
- Den tredje løsningen er å tilegne tekstene som mangler verdier, en form for gjennomsnittlig eller typisk verdi, for eksempel medianen av verdiene for resten av tekstene. Dette kan sees på som en konservativ tilnærming, ettersom det vil redusere variansen i materialet.[8]

De to siste alternativene kan evalueres ved simulering. Simuleringen kan gjøres enten på et utvalg reelle variabler eller på en teoretisk fordeling, men en teoretisk fordeling forutsetter valg av både distribusjonsparametre og egenskaper relatert til forholdet mellom håndtekster og tastetekster. Ettersom dette er egenskaper jeg forsøker å avdekke gjennom analysene, velger jeg en mer preteoretisk tilnærming ved å simulere på en variabel i materialet, nemlig subklaususer med elidert subjunksjon, både nominale og relative.

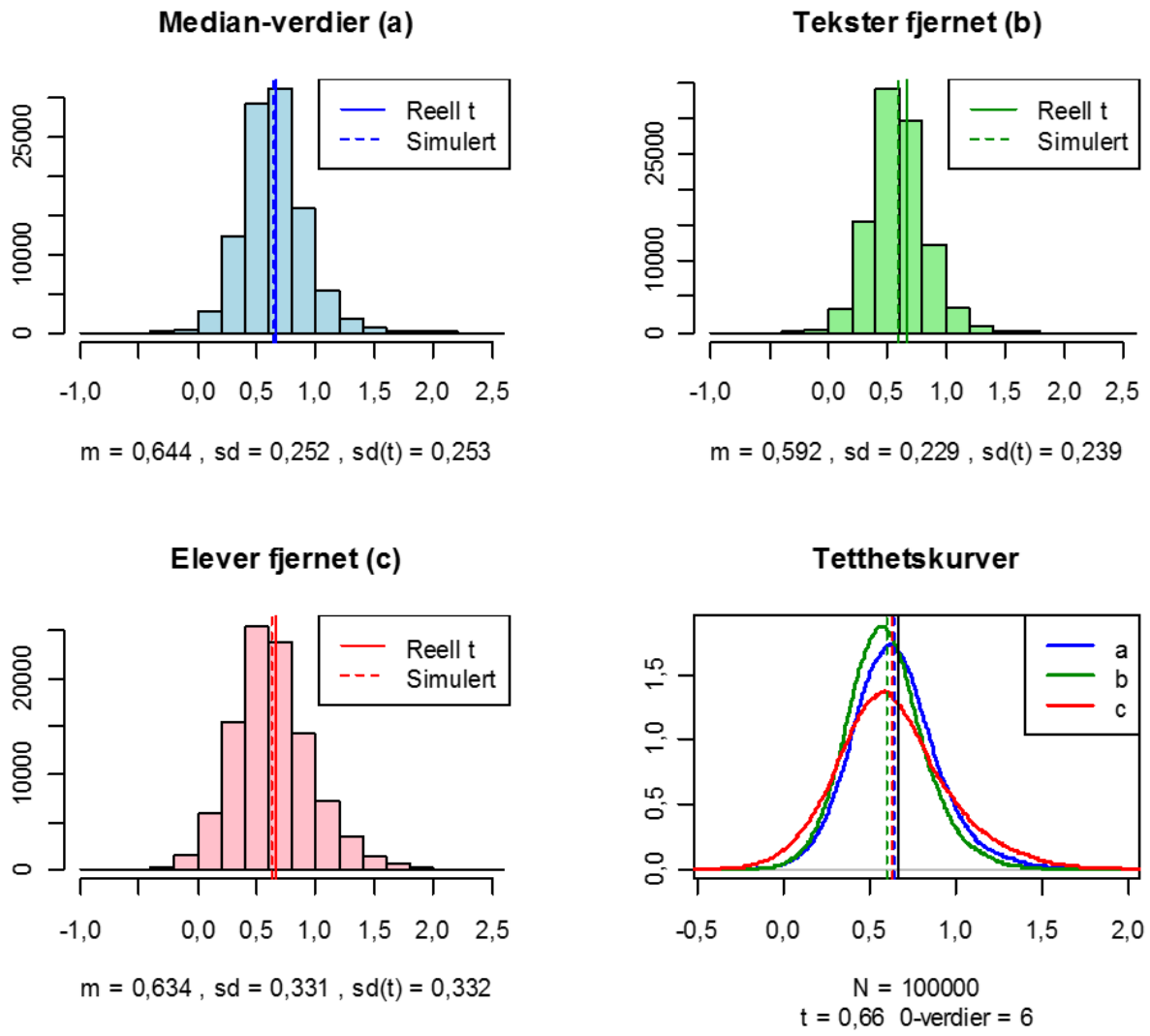
Selve simuleringen ble gjennomført på følgende måte. Jeg trakk et lite antall tilfeldige tekster fra korpuset og tilegnet disse tekstene verdien 0 for antall potensielle forekomster. Deretter jeg regnet jeg ut logit-verdiene og utførte en t-test på de resulterende verdiene med skriveverktøy som faktor. Logit-verdiene regnet jeg ut på tre forskjellige måter:

a) Ved å sette inn median-verdien for variabelen i de aktuelle tekstene og bruke disse verdiene i logit-transformasjonen. Medianverdiene ble beregnet på grunnlag av de gjenværende tekstene i utvalget.
b) Ved å fjerne de aktuelle *tekstene* og deretter regne ut logit på grunnlag av de gjenværende tekstene. Ettersom flere elever dermed bare har én tekst med i analysen, benyttet jeg i dette tilfellet en uparet t-test.
c) Ved å fjerne de *elevene* som har skrevet de aktuelle tekstene, før utregning av logit. På denne måten kunne jeg fortsatt benytte parede t-tester, men utvalget ble ytterligere redusert i størrelse.

De to siste alternativene har dessuten den ulempen at de medfører ubalanserte delutvalg, noe som svekker robustheten ved t-test og variansanalyse (7.2.3.1).

Simuleringen utførte jeg på frafall av 1, 2, 3, 6, 12 og 24 tekster med 100 000 repetisjoner for hver frafallsstørrelse. Deretter beregnet jeg middelverdien og standardavviket for t-verdiene for hver frafallsstørrelse. Resultatene for frafall av størrelse 6 er oppsummert i figur 7-5.

[8] Den intuitivt enkleste løsningen, å tilegne variabelen verdien 0,5, vil derimot være lite valid, ettersom denne verdien ofte vil ligge utenfor eller i utkanten av verdiområdet i resten av utvalget.



Figur 7-5: Simulering av nøyaktigheten i ulike beregningsmåter for logit-transformasjoner på variabelen subjunksjonsløse nominal- og relativklaususer med frafall. Antall frafall er 6.

Det tilfeldige utvalget er gjort med en enkel funksjon for tilfeldig tallgenerering fra en uniform fordeling.

(82) `nv <- ceiling(runif(n2, max=v1))`

Dette tilsvarer selektering med tilbakelegging og medfører at enkelttekster og enkeltelever kan være trukket ut flere ganger i samme utvalg, slik at utvalgene ikke er garantert å være akkurat like store i hvert forsøk. Dette var åpenbart et uheldig valg og et unødvendig, forstyrrende element i forsøket, men det har neppe stor betydning for resultatet, særlig ikke for små frafallsstørrelser, som er det som er mest relevant for studien.

Den reelle t-verdien fra en paret t-test på den logit-transformerte variabelen med skriveverktøy som prediktor er $t \approx 0{,}660$, $df = 59$. Figuren viser at middelverdien for de simulerte t-verdiene ligger noe lavere for alle metodene, noe som er naturlig ettersom utvalget er redusert. Avviket mellom middelverdien og den reelle t-verdien er størst for metode *b*, med grønn markering i figuren. Dessuten kommer det frem at standardavviket for

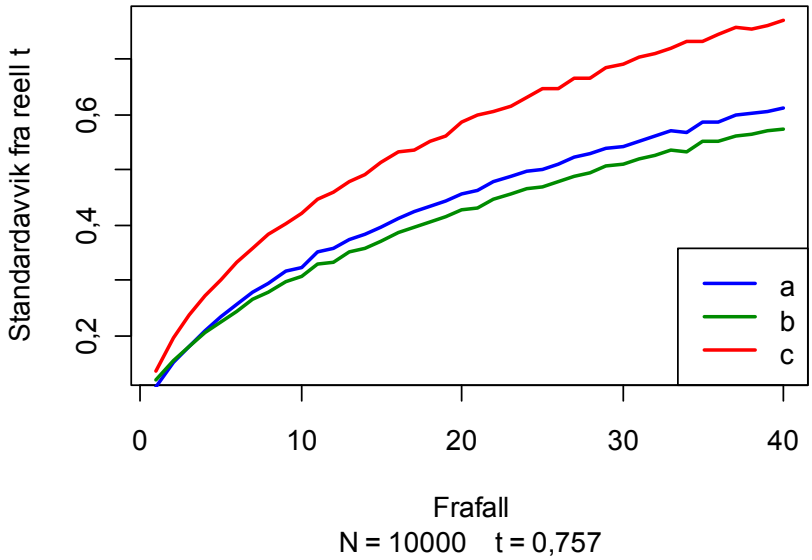
de simulerte t-verdiene er vesentlig større for metode *c* enn for de to andre; usikkerheten ved ett enkelt resultat vil altså være større for metode *c* (rød i figuren) enn for metode *a* og *b*. Dette skulle peke ut metode *a*, altså innsatte medianverdier, som den beste metoden for å kompensere for manglende potensielle omgivelser.

Det er likevel et spørsmål hvilken rolle standardavviket spiller, for selv om forskjellen i standardavvik er størst mellom metode *c* og de to andre, er også standardavviket noe mindre for metode *b* enn metode *a*. Imidlertid er det ikke standardavviket fra *middelverdien* som er det mest relevante målet på usikkerhet i dette tilfellet, men avviket fra den reelle t-verdien i hele utvalget. Hvis vi bruker formelen for standardavvik, men bytter ut middelverdien med den reelle t-verdien 0,660, finner vi at avviket fra reell t-verdi er minst for metode b, til tross for at avviket mellom reell t-verdi og middelverdien for simuleringene er størst for denne metoden:

$$sd_t(a) = \sqrt{\frac{\sum(t_a - t)^2}{n - 1}}$$

Dette betyr at metode *b* gir de mest nøyaktige resultatene, selv om middelverdien har det største avviket. Simuleringer med frafall av størrelse 3, 12 og 24 viser den samme tendensen. For frafall av enkelttekster gir metode *a* den beste tilnærmingen, mens frafall av størrelse 2 gir liten forskjell mellom de to. 30 gjentatte forsøksrunder med 2 frafall gir marginalt bedre resultat for metode *b* ($p < 0{,}01$ med paret t-test), men forskjellen er så liten at den ikke har noen praktisk betydning.

Det er flere forbehold forbundet med slike simuleringer. For det første vil det alltid være en del variasjon i resultatene, men med 100 000 repetisjoner synes resultatet å være ganske stabilt. Resultatene varierer dessuten med frafallsstørrelse, som vi har sett. Figur 7-6 viser hvordan avviket fra *t* øker med økende frafall for de tre metodene. Figuren demonstrerer hvordan kvalitetsmålet krysser for metode *a* og *b* rundt x = 2, og at metode *c* er vesentlig svakere i hele området.



Figur 7-6: Avvik fra t med frafall av økende størrelse med ulike metoder for utregning av logit. Avvik utregnet som standardavvik fra reell t-verdi. Se forklaring ovenfor.

Det kan være sammenheng mellom antall potensielle omgivelser og variabelverdiene som gjør at det simulerte frafallet burde vektes mot antall potensielle omgivelser. For eksempel kan det tenkes at elever som bruker mange nominalklaususer, har større tendens til å elidere subjunksjonen. I dette eksemplet er det ingen slike tendenser; det er ingen korrelasjon mellom andel elisjoner og antall potensielle omgivelser (rho ≈ 0,084, p ≈ 0,36). Imidlertid viser figur 7-4 ovenfor en viss tendens til større *variasjon* for lave x-verdier. Dette er som ventet, i og med at tilfeldig variasjon i forholdstall vil gjøre større utslag når x-verdiene er lave, og det illustrerer poenget med å ta mindre hensyn til forholdstall med lave absolutte verdier, slik en *mixed modelling*-tilnærming ville gjøre.[9] Dette argumentet taler for å benytte estimerte verdier fra en lineær modell i stedet for median-verdier for på den måten å oppnå et bedre estimat av hva verdien ville ha vært for en tekst dersom den hadde vært lang nok til å inneholde potensielle omgivelser. Imidlertid øker en slik tilnærming risikoen for å overtilpasse modellen, slik at den bidrar til å forsterke effekten av tendenser som bare skriver seg fra tilfeldige utslag i utvalg og ikke eksisterer i populasjonen.

Et vesentlig forbehold med dette eksperimentet er at simuleringen er utført for bare én variabel, og det er ikke sikkert andre variabler har de samme egenskapene. Blant annet kan egenskapene tenkes å variere med variabelens t-verdi, standardavvik eller andre egenskaper ved variabelens distribusjon.

Forsøkene peker i retning av at metoden med å erstatte manglende verdier med medianverdier bare er marginalt dårligere enn å fjerne tekstene med manglende verdier fra analysen, slik det ellers er vanlig å gjøre. Forsøkene viser også at fremgangsmåten er vesentlig bedre enn å fjerne begge tekstene for elever som mangler verdier i minst 1 av

[9] En alternativ løsning er å dempe utslagene på disse verdiene ved å trekke dem mot medianen.

tekstene, noe som vil være nødvendig for å utføre den typen lineær modellering av *forskjeller* som jeg har skissert tidligere i 7.3.2. Fjerning av enkelttekster (metode *b*) er ikke kompatibel med den analysedesignen med anova-modellering av differanseverdier som er beskrevet ovenfor, og det er derfor metode *a*, innsetting av medianverdier, jeg benytter i analysene der logit-transformasjoner er nødvendig. R-koden for den logit-funksjonen som er beskrevet over, står i appendiks E3.

7.3.2.4 Repeterte observasjoner

Variansanalyse og regresjonsanalyse forutsetter uavhengige observasjoner, men kan enkelt tilpasses en analyse med repeterte observasjoner. Imidlertid er repetisjonsstrukturen i denne studien svært enkel, ettersom hvert individ observeres bare to ganger. Dessuten er hypotesen som skal testes, utelukkende knyttet til forskjellen mellom de to observasjonene. På grunnlag av disse to egenskapene ved forsøket er det mulig å modellere forsøkene som uavhengige *observasjoner av differanser* og benytte ordinær multifaktoriell variansanalyse på kjønn, ferdighet, total tekstlengde og forskjell i tekstlengde.

I variablene som reelt er forholdstall og har slike egenskaper, kan responsvariabelen være differansen av de logit-transformerte variablene. I noen tilfeller ligger forholdstallenes verdiomfang såpass sentralt mellom 0 og 1, altså der forholdstallene har temmelig lineære egenskaper, at det i praksis spiller liten rolle om analysen gjøres på reelle eller logit-transformerte verdier. I slike tilfeller brukes utransformerte verdier. I lognormale variabler brukes differanse av logaritmetransformerte verdier.

7.3.3 Oppsummering

Som vist i de foregående avsnittene er normalmodellen for analyse av enkeltvariabler i denne avhandlingen en multifaktoriell anova-analyse med uavhengige (ikke-repeterte) utvalg. Dette oppnås ved å preparere data på følgende måter:

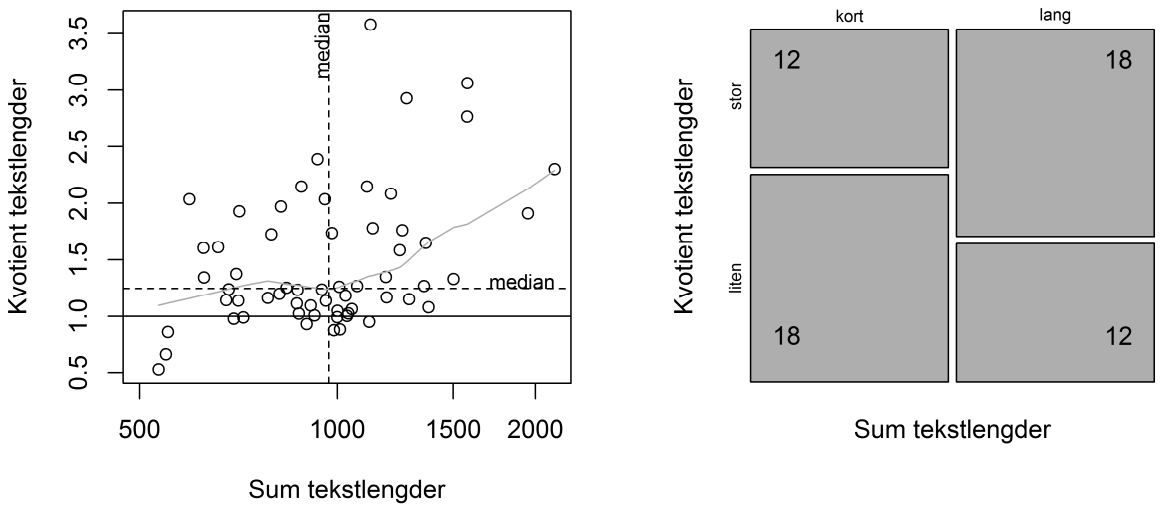
- De to kontinuerlige prediktorene dikotomiseres ved medianen.
- Responsvariabelen er normalt differansen mellom frekvensverdier i tastetekster og i håndtekster. Responsvektoren blir dermed av lengde 60, og alle de 60 variabelverdiene blir å regne som prinsipielt uavhengige observasjoner.
- For variabler som er lognormale, er responsvariabelen normalt differansen av logaritmetransformerte frekvensverdier.
- For variabler som av natur er forholdstall, som har begrenset utstrekning mellom 0 og 1 og ikke-konstant varians, er responsvariabelen normalt differansen av de logit-transformerte forholdstallene.
- Interaksjonen mellom prediktorene er begrenset til 2 nivåer.

7.3.3.1 Prediktorer

De fleste analysene vil dermed ta utgangspunkt i fire ganske uavhengige, dikotome prediktorer, som alle er knyttet til egenskaper ved eleven:

prediktor	verdi 1	verdi 2	beskrivelse
kjønn	G	J	Elevens kjønn, gutt eller jente
ferdighet	M	S	Elevens skriveferdighet, middels eller sterk
lengde	kort	lang	Elevens to teksters totale tekstlengde summert
forskjell	liten	stor	Tastetekstens lengde dividert på håndtekstens lengde

Som vist i 8.4.2 og 7.3.2.2 er det korrelasjon mellom elevenes totale tekstlengde og forholdet mellom tekstlengder. Også mellom de dikotomiserte parametrene er det noe sammenheng, som vist i figur 7-7. Det medfører behov for en viss varsomhet i tolkningen av de to parametrene i variansanalyse, særlig i modeller der de interagerer. Tabell 7-1 viser interaksjoner mellom alle de 4 prediktorene og viser også hvor små cellene er når hele utvalget av elever blir delt på fire prediktorer. (Se også 8.1 og 8.4.2 om egenskaper til elever og tekster.) Det er dermed ikke overraskende om enkelte variabler viser samvariasjon for de to parametrene. Imidlertid er det mange variabler som *ikke* viser slik samvariasjon, men bare gir utslag for den ene parameteren, noe som viser at de to parametrene i hvert fall i en viss grad er knyttet til ulike egenskaper ved elevene.



Figur 7-7: Forholdet mellom sum av tekstlengder og forholdet mellom tekstlengder. Spredningsdiagram til venstre og mosaikkdiagram til høyre.

Tabell 7-1: Interaksjonen mellom de fire prediktorene

		Korte tester			Lange tekster			Sum
		Middels	Sterke	Sum	Middels	Sterke	Sum	Sum
Liten forskjell	Gutter	6	3	9	2	4	6	15
	Jenter	5	4	9	2	4	6	15
	Sum	11	7	18	4	8	12	30
Stor forskjell	Gutter	5	2	7	2	6	8	15
	Jenter	3	2	5	5	5	10	15
	Sum	8	4	12	7	11	18	30
	Sum	19	11	30	11	19	30	60

7.4 Annet

7.4.1 Om målestokk

I forberedelsene til de statistiske analysene i denne studien har jeg utført korpussøk som trekker ut antall forekomster av de aktuelle leksikosyntaktiske fenomenene i hver tekst. Antall forekomster korrelerer normalt sterkt med tekstlengde, og vi er vanligvis mer interessert i hvilken *frekvens* et fenomen opptrer med enn *antall ganger* det opptrer i en tekst. Den grunnleggende definisjonen av frekvens er antall forekomster av noe som forekommer periodisk, per en tidsenhet. Men frekvensbegrepet kan enkelt overføres til andre måleenheter enn tid, for eksempel per lengdeenhet eller per person i en populasjon av personer. Dermed oppstår spørsmålet om hva som er den naturlige eller mest valide målestokken for leksikosyntaktiske fenomener.

Tradisjonelt har det vært vanlig å bruke antall løpeord i teksten som målestokk for språklige variabler i korpuslingvistikk, ofte multiplisert med 1000 for å oppnå mer lettleste verdier. Dette er metoden for eksempel Biber (1988) bruker. Tidligere i korpuslingvistikkens historie var dette en naturlig målestokk, ettersom den ikke krevde lingvistisk analyse av tekstene; mange tidlige arbeider var basert på søkealgoritmer som var nærmest helt blinde for lingvistiske trekk.

Etter hvert som korpuslingvistikken tok i bruk mer avanserte verktøy, åpnet det seg teoretisk mer interessante muligheter for valg av målestokk. Å regne antall forekomster per antall potensielle omgivelser ble dermed et prinsipp som det var naturlig å vurdere. For visse typer språklige variabler har dette åpenbart noe for seg; for eksempel kan det være naturlig å regne frekvens av presens per antall finitte verbaler. Det samme kan for eksempel gjelde t-enheter med kort forfelt (11.3.2) i denne avhandlingen, der en naturlig målestokk kan være t-enheten, og variabelen dermed blir *andelen* t-enheter som har kort forfelt. I andre tilfeller er det ikke så enkelt å avgjøre hva som er en potensiell omgivelse. For eksempel må subklaususer forankres i t-enheter, men t-enheten er ikke en potensiell omgivelse for subklaususen på samme måte, i og med at flere subklaususer kan forankres i samme t-enhet. Jeg har likevel valgt t-enheten som målestokk for subklaususfrekvens i denne studien. I visse tilfeller er det risiko for interaksjon mellom det fenomenet vi ønsker å måle, og den naturlige målestokken. Et typisk eksempel er attributive adjektiver, som – nesten utelukkende – må forankres i substantiver. Hvis imidlertid frekvensen av substantiver i teksten er lavt, kan frekvensen av attributive adjektiver målt på denne måten bli uforholdsmessig høyt i forhold til *antallet* attributive adjektiver i teksten. Å velge antall løpeord i teksten som målestokk kan dermed medvirke til at det ikke oppstår uønskede interaksjoner i variablene. En slik preteoretisk målestokk kan gjøre det mer meningsfylt å sammenligne ulike variabler i samme studie, og den kan også bidra til at det er enklere å sammenligne resultater på tvers av studier.

Valg av målestokk vil uansett måtte forankres i det konkrete forskningsspørsmålet man faktisk forsøker å besvare, så dermed er det ingen absolutte svar på dette spørsmålet. Generelt mener jeg imidlertid at det kan være gunstig å velge antall løpeord i teksten som

målestokk dersom forskningsspørsmålet ikke gir spesifikke grunner for å velge noe annet. Jeg har valgt andre målestokker for flere av variablene i denne studien; se spesielt diskusjonen i 11.1.3.6 om korte subklaususer, 11.3.3.4 om attributive adjektiver og 11.2 om antall ledd per klausus.

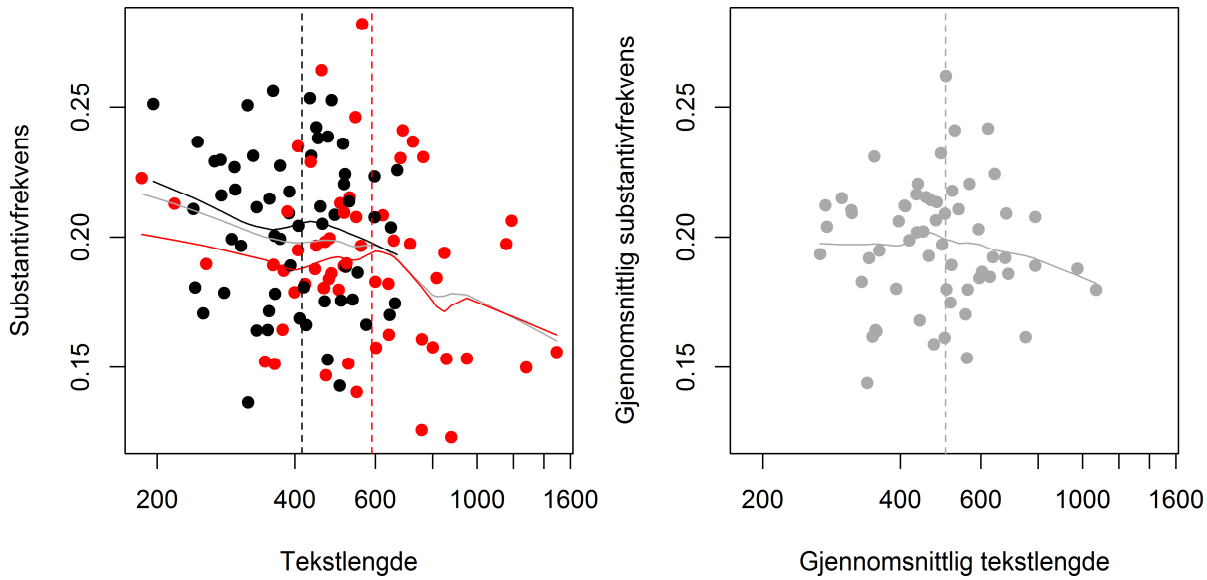
7.4.2 Korrelasjon med tekstlengde

I analysen av enkelte av de leksikosyntaktiske variablene er det relevant å undersøke korrelasjonen mellom variabelen og tekstlengde. Tekstlengden som er brukt til disse korrelasjonsanalysene, inkluderer alle ord som er skrevet av eleven, inkludert ord som står i overskrifter, fragmenter og andre steder som ikke er innenfor en t-enhet. Dessuten er antallet basert på antall grafiske ord og ikke på flerordsleksemer.

Ettersom de fleste analysene i studien baserer seg på en ordvariabel som inkluderer flerordsleksemer (8.4.1 og 9.1.1) og dessuten utelukker overskrifter, underskrifter og fragmenter (8.4.1), kan dette virke lite prinsipielt. Det er imidlertid et valg som er gjort på grunnlag av at formålet med denne delen av analysen er å studere sammenhenger mellom variabler og omfanget av produksjonen. I så fall er det naturlig å inkludere alt materialet, og ikke bare det som blir analysert av taggeren. Den enkleste måten å gjøre dette på uten manuell telling, var ved å bruke CG1, og da ble også det grafiske ordet en konsekvens (8.4.1). Å bruke det grafiske ordet har også den fordelen at det er fullstendig teoriuavhengig og dermed i større grad støtter sammenligning mellom ulike studier.

Korrelasjonsanalysene med tekstlengde er dessuten gjort på tekstnivå og ikke på elevnivå, altså med n = 120. Det innebærer at observasjonene ikke er uavhengige, men parvis avhengige, og som et resultat blir t-verdiene i disse analysene kunstig høye og p-verdiene kunstig lave. Størrelsen på dette avviket er vanskelig å fastslå og avhenger av korrelasjonen mellom håndtekster og tastetekster for den enkelte variabel. Også Shapiro-Wilk-testen er i sammenheng med korrelasjonsanalysene utført på de 120 parvis avhengige observasjonene. Selv om dette gjør korrelasjonsanalysene ugyldige som signifikanstester, gir korrelasjonskoeffisientene likevel et relevant inntrykk av korrelasjonen mellom den aktuelle variabelen og tekstlengden.

I andre delanalyser, som for eksempel når jeg sammenligner variabler etter elevens kjønn, har jeg basert analysene på elevenes middelverdier ved å summere elevens to tekster og dividere med 2 for å oppnå en sammenligning med større validitet. Når det gjelder korrelasjon med tekstlengde, virker en slik fremgangsmåte derimot risikabel, ettersom det er uklart hva som skjer med forholdet mellom en språklig variabel og tekstlengden når begge variablene reduseres til gjennomsnittsverdier, og mye av variasjonen i tekstlengde ville bli borte ved å bruke gjennomsnittsverdiene. Figur 7-8 illustrerer dette med et eksempel, nemlig korrelasjonen mellom substantivfrekvens og tekstlengde, til venstre for hver tekst og til høyre for elevenes gjennomsnittsverdier. Figuren viser tydelig hvordan spredningen reduseres når man bruker gjennomsnittsverdier.



Figur 7-8: Korrelasjon mellom substantivfrekvens og tekstlengde, til venstre målt per tekst, til høyre målt med elevenes gjennomsnittsverdier. Håndtekster er markert med svarte punkter, tastetekster med røde. Grå punkter markerer gjennomsnittsverdier for eleven.

7.4.3 Telledata

Jeg bruker i liten grad statistiske tester på telledata i krysstabeller i denne avhandlingen, men i de tilfellene det er aktuelt, bruker jeg Fishers eksakte test (Crawley, 2007, s. 309) med funksjonen `fisher.test` i R. Denne testen kan brukes på todimensjonale krysstabeller av størrelse 2x2 eller større. Et viktig premiss for denne testen og andre tester på krysstabeller, er premisset om uavhengige observasjoner. Det vil i praksis si at et individ i undersøkelsen skal være representert bare én gang i krysstabellen. Som nevnt i 7.2.2.1 er dette noe det syndes mye mot i korpuslingvistikken.

Som effektmål bruker jeg i disse tilfellene *odds ratio* (Levshina, 2015, s. 208), eventuelt kombinert med Cramérs V (Howell, 2007, s. 165).

7.4.4 Noen kommentarer om diagrammer

I en del spredningsdiagrammer i avhandlingen er det tegnet inn (rette) regresjonslinjer, gjerne ved hjelp av R-funksjonen `lm`. I noen tilfeller er det hensiktsmessig å fremheve mangelen på linearitet i regresjonen, og da har jeg i stedet for en rett regresjonslinje tegnet inn en regresjonskurve. Disse kurvene er tegnet med R-funksjonen `lowess`, som er fra pakken `stats`, som er en del av standardinstallasjonen av R (R Core Team, 2016). Funksjonen benytter lokalt vektet polynomiell regresjon, ifølge hjelpeteksten. Jeg har altså ikke foretatt noen egen ikke-lineær regresjonsanalyse for å tegne disse kurvene. Jeg bruker `lowess` også til å glatte andre typer kurver i noen diagrammer.

I boksdiagrammene er de fleste boksene i avhandlingen tegnet med innstillingen `notch=T`. Denne innstillingen tegner et "hakk" i boksen med utstrekning som tilsvarer

$$\frac{\overline{+}\ 1{,}58\ IQR}{\sqrt{N}}$$

der IQR (*inter-quartile range*) står for utstrekningen av det interkvartile verdiområdet, altså det området som omfatter de midterste 50 % av verdiene. Størrelsen på hakket representerer et overslag for et 95 % konfidensinterval for medianen, noe som medfører at ikke-overlappende hakk i to bokser gir en indikasjon på en signifikant forskjell for $\alpha = 0{,}05$. Der boksene representerer parede utvalg, slik de gjør i en del tilfeller i denne avhandlingen, mister hakkene noe av sin intuitive verdi.

Analyse

De neste fem kapitlene presenterer analyser av materialet. Det første kapitlet gir en oversikt over egenskapene til elever og tekster. De neste fire kapitlene inneholder statistiske analyser av språklige trekk i materialet basert på leksikalske og syntaktiske variabler. Disse kapitlene inneholder diskusjon og presentasjon av både metoder og resultater. Det første kapitlet behandler variabler knyttet til informasjonell tetthet. Det neste kapitlet handler om leksikalsk variasjon. Deretter følger et kapittel om syntaktiske egenskaper. Til slutt kommer et kapittel som søker å sammenfatte bildet fra de tre foregående kapitlene i en enhetlig analyse foretatt med en statistisk teknikk kalt prinsipalkomponentanalyse.

8 Elevene og tekstene

Dette kapitlet presenterer først egenskaper til de 60 elevene i utvalget, hovedsakelig fremkommet gjennom spørreskjema (6.1). Deretter presenteres noen sentrale egenskaper ved tekstene, som grunnlag for metodevalg og fortolkning av analyseresultatene i kapittel 9 – 12.

8.1 Kjønn og ferdigheter

Resultatet av utvalgsmetoden som er beskrevet i 6.1, er ikke et tilfeldig, representativt utvalg av VG1-elever i norsk videregående skole. Utvalget avviker fra et tilfeldig, representativt utvalg på flere måter, og generaliseringer fra statistiske hypotesetester på materialet må derfor gjøres med en viss forsiktighet.

For det første er utvalget begrenset av at alle elevene var elever på samme videregående skole i et tettsted på indre Østlandet. Mange av elevene har dermed den samme norsklæreren, og de fire norsklærerne samarbeider ganske sikkert i team og er inspirert og påvirket av hverandres ideer og undervisningsopplegg. Demografisk representerer de åpenbart ikke hele befolkningen av VG1-elever i Norge.

For det andre er utvalget begrenset av utvalgskriteriene som er beskrevet i kapittel 6:

- Program for studiespesialisering
- Standpunktkarakter 3 eller bedre i norsk hovedmål fra ungdomsskolen
- Norsk som morsmål, eventuelt i kombinasjon med et annet morsmål
- Bokmål som hovedmål
- Bra eller ganske bra egenrapporterte ferdigheter i Word
- Synes ikke noen av tekstbehandlingsoppgavene er vanskelige (se 6.1)

Dette er begrensninger som er gjort bevisst med tanke på å redusere variasjonen blant elevenes egenskaper ved å redusere antall parametre som kan tenkes å ha sammenheng med slike egenskaper. Den populasjonen som utvalget er hentet fra, er dermed også gjenstand for den samme typen begrensninger, og utvalget kan heller ikke på dette grunnlag sies å representere hele befolkningen av VG1-elever. Generaliseringer av funn i utvalget gjelder dermed prinsipielt bare til en populasjon med de samme parametre som er beskrevet over.

For det tredje er utvalget *balansert* mellom parametrene kjønn og skriveferdighet, slik det er beskrevet i 6.1. Utvalget er konstruert slik at det inneholder like mange gutter som jenter og like mange sterke som middels elever, og dessuten like mange sterke gutter og jenter og like mange middels gutter og jenter. Utvalget kan dermed deles i 4 like store delutvalg basert på de to parametrene kjønn og skriveferdighet, og det er denne egenskapen som gjør at jeg kaller det balansert.

Tabell 8-1: Balansert utvalg av elever over parametrene kjønn og skriveferdighet

	Middels	**Sterk**
Gutter	15	15
Jenter	15	15

Det er dermed ingen interaksjon mellom de to parametrene, i motsetning til hva som var tilfellet blant det utvalget av elever som meldte seg til prosjektet, der sterke jenter var i et klart overtall. Utvalgets fordeling av kjønn og ferdighet er med andre ord ganske sikkert ikke representativt for skolens populasjon, og heller ikke for populasjonen i Norge. Vagle (2005a) viser at det er klar sammenheng mellom kjønn og norskkarakter i KAL-materialet, og ifølge Statistisk Sentralbyrås utdanningsstatistikk (2016) hadde jenter i gjennomsnitt 4,2 som standpunktkarakter i norsk, mens guttene hadde 3,5.

Å balansere utvalget er gjort bevisst med tanke på å øke robustheten i analysene der variablene er skjevfordelt (7.2.3.1). Det medfører imidlertid at generaliseringer av funn som gjelder hele populasjonen, må gjøres med en viss forsiktighet, ettersom sterke jenter er underrepresentert i utvalget. Funn som dreier seg om interaksjonseffekter mellom kjønn og ferdighet, kan imidlertid generaliseres, mens funn som er knyttet bare til kjønn eller til kjønn og interaksjon med andre parametre enn ferdighet, må tolkes med tanke på det utvalget som er gjort.

Når det gjelder karakterene som skjuler seg bak klassifiseringen i *middels* og *sterke* ferdigheter, er guttene overrepresentert blant både dem som har karakteren 3, og dem som har karakteren 6. To elever har karakteren 3, og begge er gutter; fem elever har karakteren 6, og fire av dem er gutter. Det er ingen grunn til å anta annet enn at dette er utslag av tilfeldig variasjon, men skjevhetene *kan* tenkes å ha konsekvenser for enkelte av analyseresultatene der kjønn er en parameter, for eksempel ved at det er noe mer spredning i ferdigheter blant guttene enn blant jentene.

Tabell 8-2: Norskkarakterer fordelt etter kjønn

	3	**4**	**5**	**6**
Gutter	2	13	11	4
Jenter	0	15	14	1
Sum	**2**	**28**	**25**	**5**

Dessuten går det fram av tabell 8-2 at et overveldende tyngdepunkt av elever befinner seg i sjiktet av karakterene 4 og 5, noe som innebærer at forskjellen mellom kategoriene *middels* og *sterk* ikke er så stor som kriteriene kunne tyde på. I realiteten er sammenligningen mellom middels og sterke elever i all hovedsak en sammenligning av elever med karakterene 4 og 5. Dette kan være en medvirkende årsak til at parameteren skriveferdighet oftest ikke gir særlig store utslag i analysene i kapittel 9 – 11. Med en balansert fordeling mellom alle fire karakternivåer kunne man kanskje forvente tydeligere forskjeller. Med utgangspunkt i det begrensede elevmaterialet jeg hadde til rådighet, var en slik balansering imidlertid ikke mulig å oppnå.

Det er i noen tilfeller relevant å omtale enkeltelever. I slike tilfeller opplyser jeg noen ganger om elevens kjønn og ferdighet ved å skrive "G" for gutt og "J" for jente etterfulgt av karakteren, for eksempel "J5" for ei jente med karakteren 5 i norsk. I diagrammene bruker jeg ellers som regel "G" og "J" for kjønn og "M" for middels og "S" for sterk.

I spørreskjemaet er det to typer spørsmål om ferdigheter i tekstbehandlingsverktøy som er brukt som utvalgskriterier. Den ene spørsmålstypen er en gruppe av spørsmål der elevene vurderer om de synes at en konkret oppgave som for eksempel "å slette et ord i en setning" er "lett" eller "litt vanskelig". Bare elever som har svart "lett" på alle disse spørsmålene, er med i utvalget, så her er det ingen variasjon. Den andre spørsmålstypen består i et mer overordnet spørsmål om egenvurdering av ferdigheter, og utvalget er gjort slik at bare elever som selv rapporterer at de behersker tekstbehandlingsverktøyet "bra" eller "ganske bra", er med i utvalget. I utvalget har 35 elever svart "bra", mens 25 elever har svart "ganske bra".

Disse svarene er ikke likt fordelt mellom kjønn og ferdighet, men det er ingen signifikante forskjeller. Tabell 8-3 nedenfor viser en viss tendens til at guttene har en mer positiv oppfatning av egne ferdigheter enn jentene (*odds ratio* ≈ 0.51, p ≈ 0,29 med Fishers eksakte test), og det samme gjelder de sterke elevene i forhold til de middels sterke elevene (*odds ratio* ≈ 0.67, p ≈ 0,60 med Fishers eksakte test). Tabellen viser også at de sterke guttene er spesielt overrepresentert blant dem som mener at de behersker verktøyet bra, og at det blant jentene ikke er noen forskjell mellom middels og sterke.

Tabell 8-3: Selvrapportering av "bra" og "ganske bra" pc-ferdigheter, fordelt på kjønn og allmenne skriveferdigheter

	Gutter			Jenter			Middels	Sterke
	Middels	Sterke	**Sum**	Middels	Sterke	**Sum**	**Sum**	**Sum**
Bra	9	11	**20**	7	8	**15**	**16**	**19**
Ganske bra	6	4	**10**	8	7	**15**	**14**	**11**
Sum	**15**	**15**	**30**	**15**	**15**	**30**	**30**	**30**

8.2 Holdninger til skriveverktøy

Dette underkapitlet presenterer hvordan elevene ser på de to skriveverktøyene.

Kriteriene for utvalg gjør at alle elevene rapporterer at de *behersker* pc-verktøyet, men spørsmålet om hvilket verktøy de *foretrekker*, er ikke benyttet som utvalgskriterium. Den overveiende majoriteten av elevene i utvalget rapporterer at de alltid eller oftest foretrekker å skrive på tastatur når de "arbeider hjemme med norskfaget", men 7 av elevene i utvalget sier at de oftest foretrekker å skrive for hånd. Bare 2 av samtlige elever rapporterer at de alltid foretrekker å skrive for hånd, men disse to elevene var ikke aktuelle for undersøkelsen av andre årsaker; én hadde ikke norsk som morsmål, og én rapporterte at noen av tekstbehandlingsoppgavene var vanskelige, og ble utelukket på det grunnlaget. De 7 elevene som oftest foretrekker håndskriving, er fordelt over alle kombinasjoner av kjønn og ferdighet:

	Middels	Sterke
Gutter	**1**	**1**
Jenter	**3**	**2**

Blant de som rapporterer at de alltid eller oftest foretrekker å skrive på pc, er det imidlertid noen skjevheter. For det første ser det ut til å være en tendens til at de som rapporterer at de behersker tekstbehandling bra, er overrepresentert blant de som alltid foretrekker tekstbehandling. Det er ikke så overraskende, men det er ikke en signifikant skjevhet (*odds ratio* ≈ 0,51, N = 53, p ≈ 0,27 med Fishers eksakte test). Derimot er det en sterkere tendens til at flere gutter enn jenter *alltid* foretrekker tekstbehandling, og denne tendensen til kjønnsforskjeller i verktøypreferanse er signifikant med Fishers eksakte test (*odds ratio* ≈ 0,21, N = 53, p ≈ 0,012, Cramérs V ≈ 0,37[10]) blant de 53 som foretrekker pc. Effekten er sterkere dersom man også tar med de 7 elevene som foretrekker håndskriving. Middels og sterke elever er derimot likt fordelt med hensyn til hvordan de rapporterer at de behersker verktøyet.

	Alltid pc	Oftest pc	Oftest hånd
Behersker bra	16	16	3
Behersker ganske bra	7	14	4
Gutter	17	11	2
Jenter	6	19	5
Middels	11	15	4
Sterke	12	15	3

I spørreskjemaet ble elevene bedt om å begrunne hvorfor de foretrakk det foretrukne skriveverktøyet. I figurene under vises hvor mange som svarte henholdsvis "viktig", "litt viktig", "ikke viktig" og "usant" på de forskjellige mulige årsakene til å foretrekke pc. Noen av elevene misforstod intensjonen med dette spørsmålet, så også 3 av de elevene som oftest foretrakk håndskriving, har krysset av på hvorfor de foretrekker pc. Disse er ikke fjernet fra utvalget, og det er dermed 56 elever som er med i oversikten. Spørsmålet var "Hvorfor foretrekker du å skrive på pc i forbindelse med norskfaget?" Resultatene vises i tabell 8-4 nedenfor.

Tabell 8-4: "Hvorfor foretrekker du å skrive på pc i forbindelse med norskfaget?"

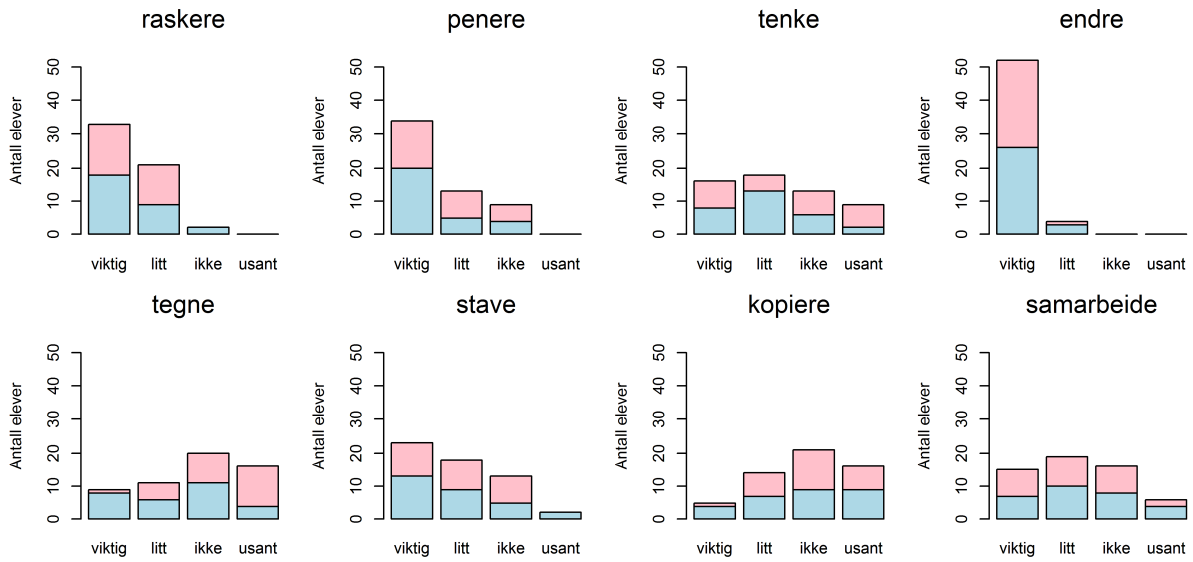
	Viktig	Litt viktig	Ikke viktig	Usant
Det går raskere	33	21	2	0
Det blir penere	34	13	9	0
Jeg tenker bedre	16	18	13	9
Det er lettere å gjøre endringer	52	4	0	0
Det er lettere å lage tegninger og figurer	9	11	20	16
Pga. stavekontrollen	23	18	13	2
Det er lettere å kopiere fra Internett	5	14	21	16
Det er lettere å samarbeide med andre elever	15	19	16	6

[10] Se (Howell, 2007, s. 165). Programkoden for Cramérs V står i E2.

Av tallene går det fram at redigeringsmuligheten er den begrunnelsen som er oppgitt som viktig desidert flest ganger, men et stort flertall mener også at raskere produksjon og penere produkt er viktig. Noen resultater er kanskje litt overraskende; at hele 16 elever mener at det *ikke* er lettere å kopiere fra Internett med pc enn for hånd, kan tyde på at noen tolker dette delspørsmålet som et spørsmål om hvorvidt de bruker pc-en til å jukse med.

Spørreskjemaet er utformet på en måte som kan forsterke elevenes positive holdninger til verktøyet, og tallene i tabell 8-4 ovenfor bør nok ikke brukes til å underbygge konklusjoner om hvorfor elever foretrekker pc over håndskriving. Men jeg synes likevel det er interessant at så mange elever finner såpass mange ulike grunner til å skrive med pc. Samtidig er det verdt å merke seg at 22 av 56 elever *ikke* svarer positivt på spørsmålet om de tenker bedre når de skriver med pc.

I figur 8-1 nedenfor er resultatene framstilt som histogrammer, og de er skilt på kjønn. Det er ingen markerte forskjeller mellom kjønnene; det tydeligste utslaget er at nesten bare gutter (8 av 9) synes det er viktig at det er *lettere* å lage tegninger og figurer på pc, mens flest jenter (12 av 16) mener at det ikke *er* lettere å lage tegninger og figurer på pc. Kanskje er dette et uttrykk for forskjeller i ferdigheter mellom kjønnene, eller eventuelt forskjeller i hvordan de oppfatter egen ferdighet.



Figur 8-1: Begrunnelser for å bruke pc til skrivearbeid i norskfaget, fordelt etter kjønn

Heller ikke om man sammenligner middels og sterke elever, er det store forskjeller mellom delutvalgene. Tydeligst er utslagene når det gjelder stavekontrollen, der flest middels elever synes å mene enten at stavekontrollen er viktig eller at den *ikke* er viktig, mens flere sterke enn middels elever (14 av 18) mener den er litt viktig. (Tallene er ikke vist her.) Ingenting tyder på at dette er annet enn tilfeldige forskjeller.

De svarene som er av størst relevans for denne undersøkelsen, er de som angår hastighet, redigering og planlegging. Alle de 56 elevene er enige om at det er enklere å redigere på pc, og de aller fleste (52) synes dette er viktig. Alle er dessuten enige om at det går raskere å

skrive på pc, og de aller fleste (54) synes dette er viktig eller litt viktig. Disse to spørsmålene er tett knyttet til hver sin delhypotese i studien. En stor majoritet, men ikke alle, mener dessuten at det er viktig eller litt viktig at produktet blir penere, noe som *kan* være knyttet til motivasjon for skrivingen, som også er en del av hypotesen. Én av elevene (en gutt med karakteren 6 i norsk) nevner dette konkret i den ene teksten i materialet når han skriver at han ikke liker å skrive for hånd fordi skriften hans "minner om små maur". Derimot mener 9 elever at de *ikke* tenker bedre når de skriver på pc, mens 13 til svarer "ikke viktig" på dette spørsmålet. Dette skulle tyde på at en betydelig andel av elevene ikke ser på pc-en som noe som fremmer planlegging av tekstene.

8.3 Bruk av pc

8.3.1 Omfang

Nesten 90 % av elevene (53 av 60) bruker pc hjemme hver dag; 1 jente bruker pc sjeldnere enn 1 dag i uka, men det er ingen andre forskjeller mellom kjønnene.

	Hver dag	1-4 dager per uke	Mindre
Gutter	27	3	0
Jenter	26	3	1
Sum	**53**	**6**	**1**

Når det gjelder hvor lenge pc-en er i bruk de dagene den er i bruk, fordeler elevene seg nesten helt jevnt mellom kategoriene "over 2 timer" og "½ til 2 timer", og det er heller ikke her forskjeller mellom kjønnene. 1 jente bruker pc-en typisk mindre enn en halv time. Dette er ikke den samme jenta som bruker pc-en sjeldnere enn hver uke. Dette betyr altså at to av jentene, elev 313 og elev 305, bruker pc relativt lite.

	Over 2 timer	½ – 2 timer	Mindre
Gutter	16	14	0
Jenter	15	14	1
Sum	**31**	**28**	**1**

Heller ikke ferdighetsnivå interagerer med disse målene for bruksmengde, og det er heller ingen interaksjoner mellom hvor ofte og hvor lenge elevene bruker pc. Derimot er det – ikke overraskende – en viss sammenheng mellom hvor lenge man sitter på pc-en, og hvor mye man spiller pc-spill (se 8.3.3).

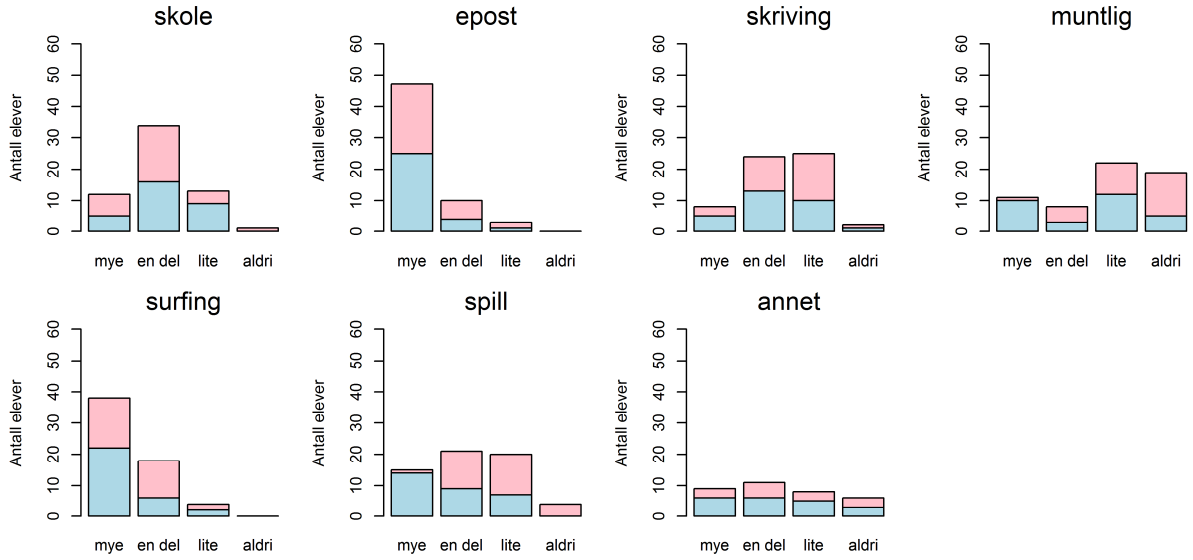
8.3.2 Aktivitet

Elevene har også svart på spørsmål om hva de bruker pc-en til hjemme. Blant aktivitetstypene er "epost, MSN, Facebook og annen skriftlig kommunikasjon" den dominerende. 47 av elevene sier at de bruker pc-en "mye" til dette.

Hva bruker du pc-en til hjemme?	Mye	En del	Lite	Aldri
Skriftlig skolearbeid	12	34	13	1
Epost, *MSN*, *Facebook*, annen skriftlig kommunikasjon	47	10	3	0

Hva bruker du pc-en til hjemme?	Mye	En del	Lite	Aldri
Annen skriving	8	24	25	2
Skype eller annen muntlig kommunikasjon	11	8	22	19
Surfing på *www*	38	18	4	0
Spill	15	21	20	4
Annet	9	11	8	6

Tallene viser at vi kan regne med at de fleste elevene er vant til å bruke pc til forskjellige skriftlige aktiviteter. 9 elever (6 gutter og 3 jenter) svarer imidlertid "lite" eller "aldri" på to av de tre mest typiske skriveaktivitetene (de tre øverste spørsmålene i tabellen), og 1 jente (elev 292) svarer "lite" eller "aldri" på alle de tre spørsmålene. Hun oppgir at hun surfer "en del" og driver med "*deviant art*" og YouTube på pc-en hjemme.



Figur 8-2: Ulike pc-aktiviteter, fordelt etter kjønn

Som det går fram av figur 8-2 ovenfor, er det få klare kjønnsforskjeller, men 14 av de 15 som sier at de spiller mye, er gutter (se mer om spillaktiviteter i 8.3.3), og 10 av de 11 som sier at de bruker pc-en mye til muntlig kommunikasjon, er også gutter. Ettersom det ikke er uvanlig å bruke *Skype* eller lignende tjenester til å snakke sammen mens man spiller over nettet, kunne man tenke seg at det er de samme guttene som spiller mye, som også kommuniserer mye muntlig, men det er ingen slik sammenheng.

Ellers er det bare små forskjeller mellom kjønnene, og det er heller ingen klare forskjeller mellom middels og sterke elever.

Tabell 8-5 nedenfor viser korrelasjoner mellom de ulike aktivitetene. Den sterkeste korrelasjonen er mellom Annen skriving og Annet, og denne sammenhengen skyldes kanskje rett og slett at begge er uspesifiserte kategorier. Ellers viser tabellen at Skriftlig kommunikasjon og Spill er de aktivitetene som korrelerer mest med de andre aktivitetene, mens Skriftlig skolearbeid ikke korrelerer med noen av de andre. At spillaktivitet korrelerer

med såpass mange av de andre aktivitetene, kan kanskje tyde på at spilling er en indikator for pc-aktivitet mer generelt. I og med at Skriftlig kommunikasjon har så dominerende score for "mye", er det vanskeligere å tolke korrelasjonene denne har med de andre kategoriene.

Tabell 8-5: Korrelasjoner mellom ulike hjemmeaktiviteter på pc. Tallene er Pearsons korrelasjonskoeffisienter. Bare koeffisienter > 0,3 er gjengitt. Ettersom distribusjonene er klart ikke-normale, må de nøyaktige verdiene tolkes med forsiktighet, men de kan likevel gi en indikasjon på sammenhenger i materialet.

	Skriftlig skole-arbeid	Skriftlig kommuni-kasjon	Annen skriving	Muntlig kommuni-kasjon	Surfing	Spill	Annet
Skriftlig skolearbeid	–	0	0	0	0	0	0
Skriftlig kommunikasjon	0	–	0	0,40	0,30	0,43	0
Annen skriving	0	0	–	0	0	0	0,49
Muntlig kommunikasjon	0	0,40	0	–	0	0,41	0
Surfing	0	0,30	0	0	–	0,34	0,33
Spill	0	0,43	0	0,41	0,34	–	0
Annet	0	0	0,49	0	0,33	0	–

8.3.3 Spill

Én av aktivitetene som elevene er spurt om i spørreskjemaet, er altså spill, og korrelasjonene i tabell 8-5 gir kanskje grunn til å tro at høy spillaktivitet er en indikator for mye og variert aktivitet på pc ellers. Jeg studerer derfor opplysningene om spilling litt nærmere.

Som for de andre aktivitetene har elevene kunnet krysse av for om de spiller "mye", "en del", "lite" eller "aldri". Bare 4 elever krysser av for "aldri". Mellom de andre alternativene er fordelingen ganske jevn, men noe færre elever har svart "mye" enn "en del" eller "lite".

Tabell 8-6: Oversikt over hvor mye gutter og jenter sier at de spiller på pc

	Mye	**En del**	**Lite**	**Aldri**
Gutter	**14**	**9**	**7**	**0**
Jenter	**1**	**12**	**13**	**4**
Sum	**15**	**21**	**20**	**4**

Tabell 8-6 ovenfor viser en tydelig tendens til at det er gutter som (rapporterer at de) spiller mest, og blant de 4 elevene som aldri spiller, er det bare jenter. Denne tendensen ser ut til å bekreftes av forskning på kjønn og dataspill (for eksempel Epstein, 2012), selv om forskjeller i rekrutteringsmetode gjør det problematisk å sammenligne resultatene mine med Epsteins. Det er ingen tendenser til interaksjon mellom elevens karakter i norsk og hvor mye det spilles.

Det er flere kommentarer som kan knyttes til disse tallene. For det første er selvfølgelig kategoriene "mye", "en del" og "lite" ganske subjektive. Gutter som har venner som spiller 4 timer per dag, synes kanskje de spiller "lite", selv om de spiller mer enn jenter som svarer "en del". På den annen side kan det tenkes at spilling er mindre sosialt akseptert blant jenter

enn blant gutter, og at dette fører til underrapportering blant jenter. Det er mye usikkerhet her, men den relativt like fordelingen mellom de tre kategoriene (15 : 21 : 20) kan tilsi at elevene har en noenlunde lik forståelse av hvor mye elever på deres alder spiller.

Å spille på pc innebærer svært ulike typer aktiviteter, avhengig av hva slags spill man spiller. Spill kan selvfølgelig kategoriseres på flere detaljnivå, men jeg har her valgt å dele dem inn i fire kategorier med tanke på egenskaper som kan ha relevans for skriving, i og med at det er skriving og tekst som er studieobjektet i denne avhandlingen:[11]

1. Spill med én eller flere brukere der aktiviteten ikke eller i svært liten grad er tekstbasert.
2. Spill av typen rollespill (*RPG – Role Playing Games*) som i ganske stor grad er tekstbasert, men som ikke innebærer noe særlig skriving fra deltagernes side.
3. Spill som innebærer samspill med andre, men der samspillet i liten grad er skrivebasert, særlig spill av typene *First Person Shooter / Third Person Shooter* og *Real-Time Strategy*. Skriftlig kommunikasjon er mulig og mer eller mindre utbredt i form av chat-funksjonalitet, men denne kommunikasjonen er ikke sentral for spillets gang.
4. Spill av typen *MMORPG – Massively multiplayer online role-playing games*, som i stor grad er basert på at brukerne kommuniserer skriftlig.

Tabell 8-7 nedenfor gir en oversikt over de spillene eller spilltypene som elevene nevner, kategorisert i de fire kategoriene. Mange elever har oppgitt navn på spill de spiller, men flere nevner også bare nettsteder. Disse nettstedene tilbyr først og fremst enkle spill i kategori 1, og jeg har derfor kategorisert nettstedene sammen med spillene i kategori 1. Noen elever oppgir også spilltyper heller enn konkrete spill. Mange elever oppgir spillets versjon i tillegg til navn, for eksempel *Call of Duty 4*. Versjonsinformasjonen er som regel irrelevant for kategoriseringen, og i de tilfellene har jeg fjernet opplysningen fra oversikten.

Tabell 8-7: Oversikt over de spill elevene nevner at de spiller, kategorisert i fire ulike typer. Se forklaring i teksten.

	1	**2**	**3**	**4**
spilltyper	kabal flashgames sjakk poker	RPG rollespill	FPS Strategispill	
nettsteder	1000spill 123spill			

[11] Det meste av informasjonen om ulike spill stammer fra personlige samtaler med Håvard Vibeto, på den tiden samtalene fant sted, doktorgradsstipendiat ved Høgskolen i Hedmark, med spill som arbeidsområde. Noe stammer også fra min sønn, som er født 1994, altså to år etter informantene i undersøkelsen, men som da jeg snakket med ham, var like gammel som informantene var ved datainnsamlingen. Jeg har også brukt informasjon fra de enkelte spills websider, samt Wikipedia.

	1	2	3	4
	sol.no-barn kongregate.com			
konkrete spill	Bomberman Castlevania Diablo Elma FIFA Football Manager Far Cry Freeride Extreme Geometry Wars Jazz Jackrabbit Mario Mercenaries Rayman Run Jibbin Sims Tetris Settlers TrackMania Worms world party	Fallout Final fantasy	Battlefield Call of Duty Command and Conquer CounterStrike Crysis Rainbow Red Alert Team Fortress Vegas Warcraft 3	Age of Conan Warhammer online World of Warcraft

10 elever har ikke spesifisert hva slags spill de spiller. Mange skriver bare en generell omtale: "forskjellig", "mye rart", "nettspill", "tidsfordriv", "alt mulig", "småspill litt her og der". Selv om "mye rart" og "alt mulig" godt kan tenkes å referere til for eksempel skytespill eller *MMORPG*, har jeg regnet med at de som ikke spesifiserer spilltype eller spillnavn, spiller de minst avanserte spillene, og jeg har derfor plassert alle disse elevene i kategori 1.

Alle elevene som har navngitt enkelte rollespill som ikke er *MMORPG*, har også nevnt minst ett *MMORPG*-spill ved navn. Jeg regner derfor med at de fleste ungdommer som spiller rollespill, vanligvis også spiller *MMORPG*. Bare 1 elev har svart "Div. rpg-spill" uten å spesifisere spillnavn, og jeg har på bakgrunn av hvordan de andre elevene har svart, valgt å anta at også denne eleven spiller *MMORPG* i tillegg til andre rollespill. Jeg kunne derfor slå sammen kategori 2 og kategori 4.

Resultatet blir derfor bare 3 kategorier av spill:

1. Spill som ikke fordrer noen skriving, kalt "diverse" i tabell 8-8 nedenfor,
3. Spill av typen skytespill eller strategispill, som kan inkludere noe enkel chat-skriving,
4. *MMORPG*-spill, som er basert på omfattende skriving.

På grunnlag av dette har jeg også kategorisert elevene i 4 kategorier, som vist i tabell 8-8 nedenfor, nemlig elever som spiller spill i kategori 4, elever som spiller spill i kategori 3

men ikke kategori 4, elever som spiller spill i kategori 1 men ikke 3 eller 4, og elever som ikke spiller spill.

Tabell 8-8: Sammenheng mellom hvor mye elevene spiller og hva de spiller, fordelt etter kjønn

Mengde:	**Mye**			**En del**			**Lite**			**Aldri**		
Spilltype	G	J	**Sum**	G	J	**Sum**	G	J	**Sum**	G	J	**Sum**
MMORPG	7	0	**7**	0	2	**2**	0	0	**0**	0	0	**0**
Chat	5	0	**5**	3	0	**3**	0	0	**0**	0	0	**0**
Diverse	2	1	**3**	6	10	**16**	7	13	**20**	0	0	**0**
Ingen	0	0	**0**	0	0	**0**	0	0	**0**	0	4	**4**
Sum	**14**	**1**	**15**	**9**	**12**	**21**	**7**	**13**	**20**	**0**	**4**	**4**

Som nevnt er det bare 4 elever som ikke spiller, og disse er jenter. Blant resultatene ellers er det mest påfallende:

1. Det er en klar sammenheng mellom hvor mye man spiller, og hva man spiller. Ingen som spiller "lite", spiller spill i kategoriene 3 eller 4. Blant de som spiller "mye", spiller de fleste spill i kategori 3 og 4.
2. Bare gutter spiller skyte- eller strategispill.
3. Nesten bare gutter spiller MMORPG.
4. Det er ingen sammenheng mellom skriveferdigheter og spillvaner; middels og sterke elever er jevnt fordelt i alle kategorier.

Det er flere potensielle feilkilder i denne fremstillingen. For det første sier ikke undersøkelsen noe om hvor mye hver elev spiller av hver kategori. For det andre kan det tenkes at sammenhengen mellom kvantitet og kategori skyldes at elever som spiller mye, finner det mer naturlig å spesifisere hva de spiller, og at noen av dem som ikke har spesifisert spill, dermed har havnet i feil kategori. På den annen side er nok *MMORPG* en type spill som på en annen måte enn for eksempel *Tetris* eller *Mario* fordrer mye spilling.

Tabell 8-9 nedenfor viser en viss tendens til sammenheng mellom hvor mye man spiller, og hvor lenge man sitter ved pc-en, ikke overraskende.

Tabell 8-9: Sammenheng mellom hvor mye elevene spiller, og hvor lenge de sitter ved pc-en

	Over 2 timer	**1/2 – 2 timer**	**Under 1/2 time**	**Sum**
Mye	10	5	0	**15**
En del	13	8	0	**21**
Lite	7	13	0	**20**
Aldri	1	2	1	**4**
Sum	**31**	**28**	**1**	**60**

Fremstillingen av spillaktiviteter viser klare kjønnsforskjeller som ikke framkommer for andre typer av pc-aktiviteter.

8.4 Tekster

Dette delkapitlet presenterer noen kvantitative egenskaper ved tekstene som har konsekvenser for metodevalg og metodeutvikling som jeg presenterer og drøfter i 9, 10 og 11.

8.4.1 Beregning av tekstlengde

Tekstenes lengde er interessante av flere ulike årsaker. For det første kan man regne med at tekstlengden henger sammen med både produksjonshastighet og motivasjon, og altså slik har en forbindelse til hypotesen. I tillegg spiller tekstenes lengde en rolle i mange av de språklige variablene i undersøkelsen. Dessuten er det en generell statistisk sammenheng mellom elevteksters lengde og kvalitet, slik det blant annet går fram av KAL-prosjektet (Evensen, 2003, s. 8). Hultman og Westman (1977, s. 54) finner også en sterk korrelasjon mellom lengde og bedømming, dog ikke for de høyeste karakterene. Også Östlund-Stjärnegårdh (2002, s. 76) finner sammenheng mellom lengde og bedømming, men hennes fokus er på skillet mellom godkjente og ikke-godkjente tekster og dermed ikke så relevant for denne undersøkelsen.

Jeg beregner tekstlengde som antall løpeord i teksten, men på bakgrunn av de ulike formål for å studere tekstlengde, beregner jeg den på to ulike måter. De to måtene følger naturlig fra de litt ulike parseralgoritmene i de to versjonene av taggerprogrammet, CG1 og CG3 (se 6.4.2). CG1 bruker det grafiske ordet som enhet, altså alle sammenhengende bokstaver eller tall adskilt av mellomrom eller skilletegn. CG3 legger til grunn en lemmatisering basert på en ordliste som inkluderer et lite antall flerordsleksemer. (Se 9.1.1 nedenfor for en mer detaljert diskusjon av flerordsleksemer.) Begge algoritmene ser bort fra ord i overskrift og underskrift, men CG1 regner med ord som står i klaususfragmenter, mens CG3 overser leksemer i ekte fragmenter.

Jeg bruker tallene fra CG1 som grunnlag for beregninger av hvor mye tekst elevene har produsert. Det vil si at de to prediktorvariablene knyttet til tekstlengde, total tekstlengde og forskjell i tekstlengde, er basert på grafiske ord og inkluderer ord i klaususfragmenter. Som uttrykk for produksjonsmengden er det åpenbart fornuftig å inkludere ord i fragmenter; disse ordene er jo også produsert, de former en del av teksten og har krevd kognitiv kapasitet å produsere. At de står i klaususfragmenter, er neppe noe elevene har et bevisst forhold til mens de skriver. Overskriftene og underskriftene er selvfølgelig også produsert, men de er neppe i like stor grad en integrert del av den kognitive produksjonsprosessen, og det er derfor ikke urimelig å utelate dem fra den registrerte produksjonsmengden. Å basere antall ord på det grafiske ordet og ikke på en leksemdefinisjon kan nok ikke begrunnes like godt ut fra et kognitivt argument, kanskje tvert imot. Dersom disse leksemene faktisk har status som leksemer i hjernens leksikon, er det neppe mer kognitivt krevende å hente dem fram og produsere dem enn å hente fram og produsere et enkeltordsleksem. Dette gjenspeiles av at staving av flerordleksemer er et litt flytende område i norsk rettskrivning, både i den historiske utviklingen av offisiell rettskrivning og i hvordan elever og andre skribenter forholder seg til den. Et illustrerende eksempel er at to elever i korpuset skriver \hvertfall\

(uten \i\) i stedet for \i hvert fall\, som er det riktige ifølge den offisielle rettskrivningen. Imidlertid er det en fordel for sammenlignbarhet med andre studier å basere parsingen på det grafiske ordet i henhold til den offisielle rettskrivningen, slik jeg har gjort.

Jeg bruker derimot tallene fra CG3 som grunnlag for beregning av frekvens- og lengdemål for de syntaktiske variablene, for eksempel gjennomsnittlig antall ord per klausus, ettords forfelt, antall attributive adjektiver per løpeord, og for de leksikalske variablene, for eksempel gjennomsnittlig ordlengde, leksikalsk tetthet og TTR-baserte mål. Å utelukke ekte fragmenter fra beregning av syntagmelengder er logisk og fornuftig, men når det gjelder de leksikalske variablene og enkelte andre frekvensmål, var det rett og slett bare en nødvendig konsekvens av å bruke CG3-versjonen av taggeren. Validitetsmessig er det nok litt uheldig, men konsekvensene er små. Også konsekvensene av å legge flerordsleksemer til grunn for leksikalsk statistikk og andre frekvensmål er små; jeg diskuterer dem noe mer inngående i avsnitt 9.1.1 om gjennomsnittlig ordlengde.

Som forklart i 6.4.2 ble korpusteknologien byttet ut i løpet av prosjektet, og mesteparten av søkene og analysene i undersøkelsen er utført med CG3-parseren. Dette legger de rammene for variablene som er forklart i avsnittene over, og det var dermed ikke betingelser jeg hadde noen særlig praktisk innflytelse over. Jeg mener imidlertid at rammene hovedsakelig er validitetsmessig gunstige, og at de mulige negative konsekvensene ikke er store. Jeg mener likevel det er validitetsmessig riktigst å inkludere ord i fragmenter i de variablene som skal reflektere omfanget av elevenes produksjon, og da var det nødvendig å benytte CG1-parsingen til dette, med de følger det har. For noen tekster har dette merkbare konsekvenser; den største forskjellen i tekstlengde mellom de to beregningsmåtene er 26 ord. Men for de fleste tekstene har det liten betydning; medianen for forskjellen er bare 3 ord, og for 17 tekster er det ingen forskjell.

Til grunn for parsingen ligger tekster som er korrigert for gjeldende rettskrivning. Dette er nødvendig for at parseren skal kunne gjenkjenne leksemene. Det innebærer altså at de sær- og samskrivninger som ikke er i tråd med rettskrivningen, er korrigert, og antall grafiske ord som beregnes, er derfor ikke alltid akkurat det samme som det antall grafiske ord eleven skrev.

Ifølge Halliday (se 3.1) skiller muntlig og skriftlig språkbruk seg blant annet ved at det er flere grammatiske ord i muntlig språk, mens antall leksikalske ord er mer eller mindre det samme for det samme innholdet. I lys av dette er kanskje antall leksikalske ord – eller helst antall leksikalske morfemer – et mer relevant mål for tekstlengde, ved at det bedre gjenspeiler mengden av innhold i teksten. Antall leksikalske morfemer er ikke automatisk tilgjengelig fra korpuset, men antall leksikalske ord har først og fremst den ulempen at definisjonen er prototypisk, slik jeg diskuterer i 9.2.1, noe som vanskeliggjør sammenligning med andre studier. Dessuten er korrelasjonen mellom antall ord og antall leksikalske ord i tekstene så sterk, $R \approx 0{,}978$, at valget ikke har noen særlig stor praktisk betydning.

8.4.2 Deskriptiv statistikk for tekstlengde

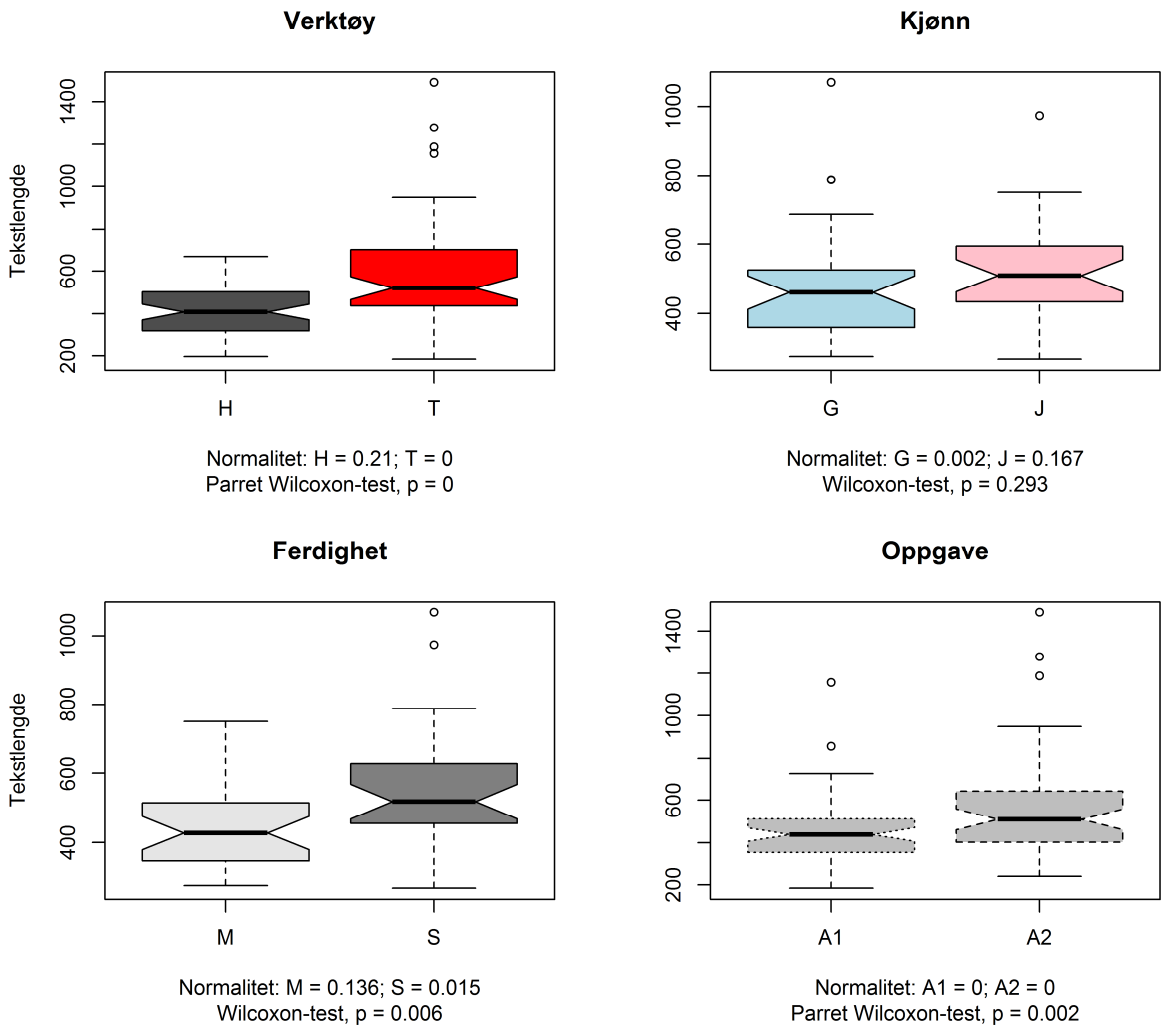
Tabell 8-10 nedenfor viser nøkkeltallene for tekstlengde målt i antall grafiske ord. Tallene er basert på CG1, altså på det grafiske ordet, der ord i fragmenter er med, men ikke ord i overskrifter og underskrifter.

Tabell 8-10: Nøkkeltall for tekstlengde i antall ord. I denne og lignende tabeller refererer *Hånd* og *Tast* til de to skriveverktøyene, *Middels* og *Sterk* til de to nivåene av skriveferdigheter, og *Gutt* og *Jente* til de to kjønnene. I overskriftsraden står *sd* for standardavvik.

	middelverdi	median	sd	min	maks
Total	**502**	**464**	**213**	**185**	**1491**
Hånd	415	408	117	196	669
Tast	589	520	250	185	1491
Middels	445	415	157	196	949
Sterk	559	506	246	185	1491
Gutt	490	450	228	218	1491
Jente	514	472	200	185	1279

Tekstene viser ganske stor variasjon i lengde, særlig tatt i betraktning at elevene har hatt bare to skoletimer til rådighet. Middelverdien er 502 ord, mens medianen er 464. Den lengste teksten er skrevet på tastatur av en G5, og den er på 1491 ord. Den korteste teksten er også skrevet på tastatur; den er skrevet av ei J5, og den er på 185 ord. Den lengste håndskrevne teksten er på 669 ord, og den er også skrevet av ei J5. Median og særlig middelverdi er vesentlig høyere blant tastetekstene enn blant håndtekstene, og 50 av de 60 elevene har skrevet lengre tastetekster enn håndtekster.

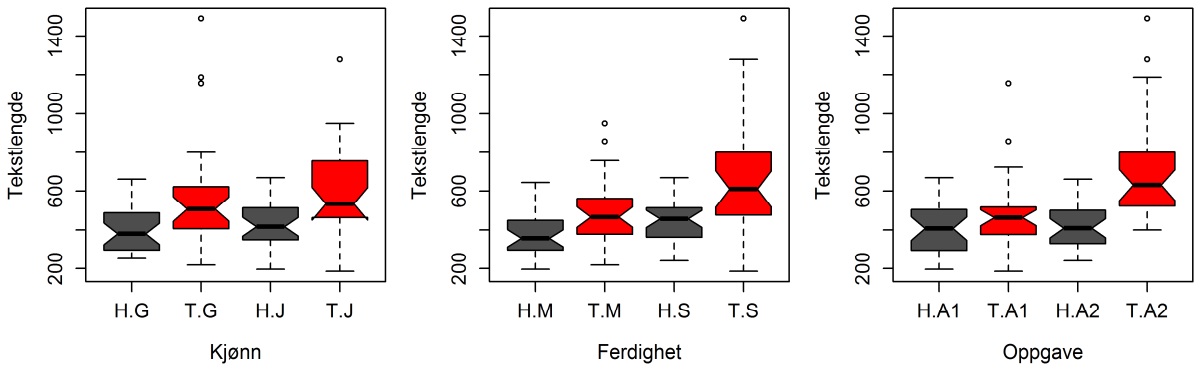
Tekstlengde er ikke normalfordelt, men de logaritmetransformerte verdiene er det. I det følgende skal vi se på sammenhenger mellom tekstlengde og de fire elevfaktorene som blir brukt i analysen av de språklige variablene. Siden jeg ikke har konkrete hypoteser knyttet til tekstlengde, rapporterer jeg ikke signifikans, men de fire faktorenes sammenheng med tekstlengde i utvalget er relevant for diskusjon og fortolkning av de språklige variablene.



Figur 8-3: Tekstlengde i antall ord fordelt etter fire parametre. I dette og lignende diagrammer er normaliteten i utvalgene testet med Shapiro-Wilks normalitetstest, og p-verdien fra Shapiro-Wilk-testen for hvert utvalg er gjengitt. På grunnlag av p-verdien er forskjellen mellom utvalgene testet med henholdsvis Welch' t-test eller Wilcoxons rangsum-test. For skriveverktøy og oppgave er det brukt en paret test, mens det for kjønn og ferdighet er brukt uparede tester på gjennomsnittsverdien av de to tekstene fra hver elev. Ettersom hypotesetestene ikke er knyttet til konkrete hypoteser i prosjektet, skal p-verdiene bare tolkes i et eksplorativt perspektiv.

Figur 8-3 viser hvordan tekstlengde varierer etter de to elevparametrene kjønn og skriveferdighet og de to tekstparametrene skriveverktøy og oppgave. Jentene i utvalget skriver noe lengre enn guttene. De sterke elevene skriver lengre enn de middels elevene, de tastede tekstene er lengre enn de håndskrevne tekstene, og "Ungdomsfylla"-tekstene er lengre enn "Bøker eller data"-tekstene. Selv om kjønnsforskjellen er relativt liten og p-verdien fra den parede Wilcoxon-testen er høyere enn vanlig signifikansnivå, kan forskjellen ha innvirkning på noen av de språklige variablene eller på tolkningen av dem.

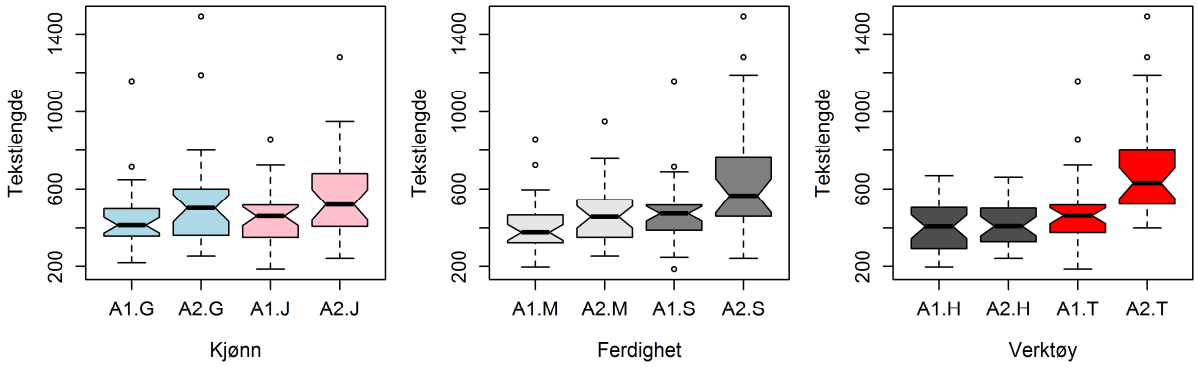
Det er også en del interaksjon mellom skriveverktøy og de andre tre parametrene, slik det går fram av figur 8-4 nedenfor.



Figur 8-4: Tekstlengde. Interaksjon mellom verktøy og kjønn, ferdighet og oppgave.

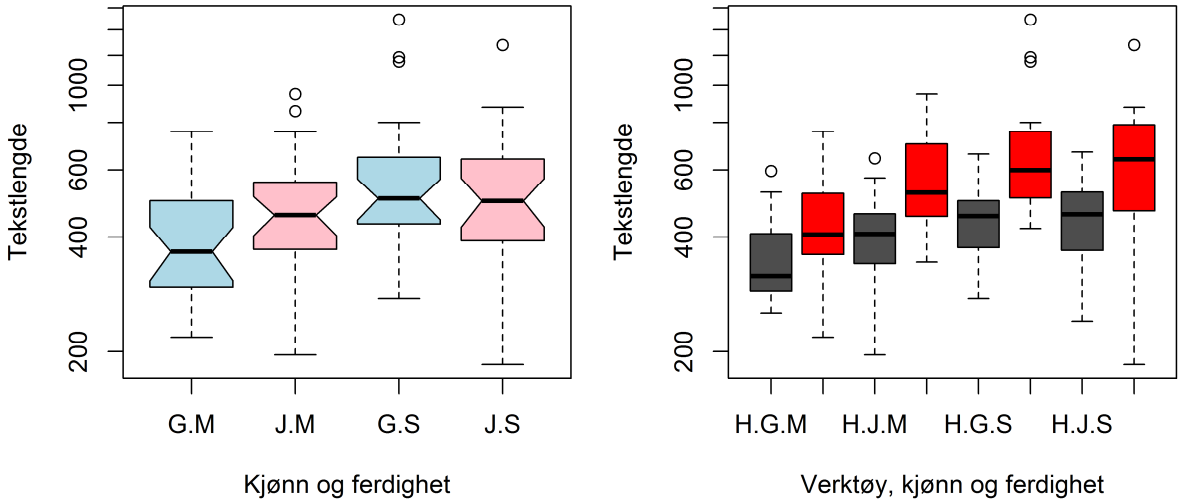
Figuren viser at det særlig er "Ungdomsfylla"-tekstene som viser stor forskjell i lengde mellom de to verktøyene. Det ser også ut til at utslaget er størst for sterke skrivere og for jenter, men forskjellen mellom kjønnene er mindre enn forskjellen mellom de to nivåene av skriveferdighet og de to oppgavene.

Siden figur 8-3 indikerer at oppgavetekst har avgjørende innvirkning på tekstlengden, er det relevant å se hvordan denne tekstparameteren interagerer med de andre parametrene. I figur 8-5 nedenfor ser vi at det særlig er sterke skrivere som skriver lenger om "Ungdomsfylla", mens oppgave ikke ser ut til å interagere med kjønn.



Figur 8-5: Tekstlengde. Interaksjon mellom oppgave og skriveverktøy, kjønn og skriveferdighet

Til slutt ser vi i figur 8-6 på interaksjon mellom kjønn og ferdighet, og vi ser at det er særlig de svakere guttene som skiller seg ut ved å skrive kortere. Tendensen til kortere tekster for gutter med middels ferdighet synes å gjelde både for hånd og på tastatur.



Figur 8-6: Tekstlengde og interaksjon mellom ferdighet og kjønn, og mellom verktøy, ferdighet og kjønn. Logaritmisk y-akse.

Noe som er verdt å merke seg, er at *all* forskjell i lengde skriver seg fra oppgaven om "Ungdomsfylla". Det kan være flere årsaker til det.

Som forklart i 6.2, skrev alle elevene "Bøker eller data"-teksten først og "Ungdomsfylla"-teksten etterpå innenfor et undervisningsopplegg om argumenterende tekster. Dette kan ha hatt flere effekter. For det første vil elevene i løpet av undervisningsopplegget ha lært mer om hvordan man skriver en argumenterende tekst, eller i hvert fall ha fått bevisstgjort tidligere tilegnet kunnskap om dette. Kanskje har de sterke elevene i større grad enn de middels elevene klart å gjøre seg nytte av dette undervisningsopplegget, slik at denne faktoren har størst effekt på de sterke elevene. Kanskje er det utbyttet fra undervisningsopplegget som er bakgrunnen for lærernes oppfatning om at "Ungdomsfylla"-oppgaven fungerte best, og at elevene her hadde lettere for å finne argumenter å bruke i teksten, selv om en del elever fortalte læreren at de ikke forsto hvilken holdning artikkelforfatteren egentlig hadde. I "Bøker eller data"-tekstene er det derimot mange elever som strever med å argumentere, og de bruker teksten sin til å slå fast at sitatet i oppgaveteksten "tar feil". Jeg hadde forventet at flere elever tok opp underliggende temaer som hvorvidt det er ønskelig eller problematisk om vi har en kjønnsdeling i mediebruk, men lærerne var ikke så overrasket over at slike betraktninger var fraværende.

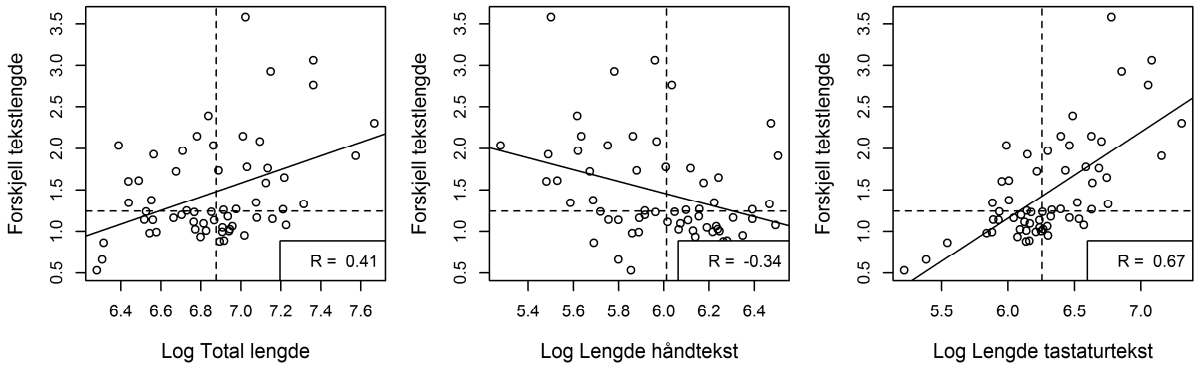
På den annen side kan man tenke seg at en del elever synes det er slitsomt å ha to skriveøkter relativt tett på hverandre (6.2) og at motivasjonen for oppgave 2 derfor ikke har vært like stor som for oppgave 1. Lengdevariabelen tyder imidlertid ikke på denne effekten, men på grunn av forskjellen i skriveverktøy kan vi vanskelig sammenligne tekstlengdene parvis for å finne tendensen. Av de 10 elevene som har skrevet *kortere* tastetekster enn håndtekster, gjelder dette imidlertid i 9 tilfeller "Bøker eller data"-teksten, altså oppgave 1.

Like interessant som hvorfor elevene har skrevet så mye lengre tastetekster enn håndtekster om "Ungdomsfylla", er hvorfor de har skrevet *like langt* med de to verktøyene om "Bøker eller data". Hvorfor genererer ikke den antatt potensielt høyere produksjonshastigheten

lengre svar på denne oppgaven? Svaret *kan* selvfølgelig ligge i at mange elever ikke har hatt argumenter å skrive om i denne teksten, og at tekstlengden reflekterer hvor langt det er mulig å skrive om emnet uten å ha utfyllende momenter.

Denne skjevheten har konsekvenser for de videre analyser. I de tilfeller vi mistenker korrelasjon mellom visse variabler og tekstlengde, må vi huske at forskjellen i tekstlengde mellom hånd- og tastetekster stort sett gjelder bare den ene oppgaven. At dette er knyttet til oppgaveparameteren, er et problem med større metodiske konsekvenser enn om det hadde angått det ene kjønnet eller det ene nivået av ferdigheter, fordi forsøkets parallelle design fortsatt ville ha vært gyldig for det andre kjønnet eller det andre nivået av ferdighet. At det dreier seg om oppgaveparameteren, betyr at vi ikke kan velge å se bort fra den ene oppgaven og deretter sammenligne tekstpar som er skrevet av samme elev; den parallelle designen ville i så fall bryte sammen.

Jeg har valgt å bruke elevegenskapene total tekstlengde og forskjell i tekstlengde som prediktorer i analysene av leksikalske og syntaktiske variabler. Valget har jeg gjort ut fra en forestilling om at nettopp disse to faktorene samspiller med de språklige variablene. De to variablene er imidlertid problematiske i den grad de er egenskaper knyttet til eleven i motsetning til de enkelte tekstene, og man kan ikke se bort fra at lengden til den enkelte hånd- eller tasteteksten kunne ha større forklaringskraft enn elevegenskapene. Figur 8-7 viser derfor korrelasjonen mellom tekstlengdeforskjell og henholdsvis totaltekstlengde, håndtekstlengde og tastetekstlengde.



Figur 8-7: Forskjell i tekstlengde. Korrelasjon med total tekstlengde (til venstre), håndtekstlengde (i midten) og tastetekstlengde (til høyre). Pearsons korrelasjonskoeffisient for hver korrelasjon er gjengitt i diagrammene.

Figuren viser at det er betydelig korrelasjon mellom tekstlengdeforskjell og total tekstlengde, slik det også går fram av krysstabellen i 7.3.2.2. Det er også en ganske svak negativ korrelasjon mellom tekstlengdeforskjell og håndtekstlengde, noe som ikke er så overraskende ettersom håndtekstlengdene står under brøkstreken når vi regner ut forholdstallet. Den sterkeste korrelasjonen er mellom tekstlengdeforskjell og tastetekstlengde; høye verdier for tekstlengdeforskjell skriver seg altså i størst grad fra tastetekstlengden.

9 Informasjonell tetthet

Dette kapitlet omhandler ulike perspektiver på informasjonell tetthet, med mest vekt på variablene ordlengde (9.1), leksikalsk tetthet (9.2) og leksikalsk spesifisitet (9.3).

9.1 Gjennomsnittlig ordlengde

Selv om gjennomsnittlig ordformlengde ikke er en særlig avansert lingvistisk variabel, kan verdien gi indikasjoner om visse tekstlige egenskaper.

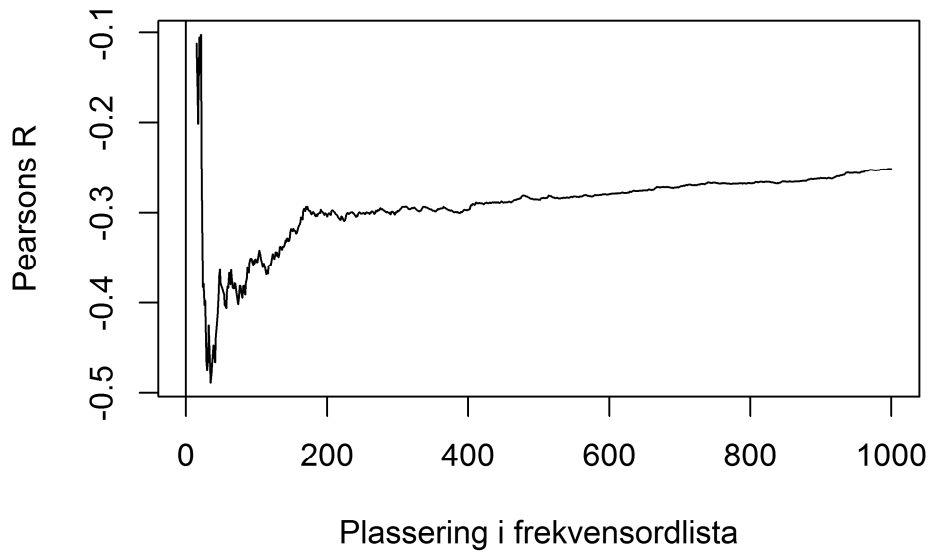
Det er en velkjent egenskap ved et språks vokabular at det er negativ korrelasjon mellom et ords frekvens og dets lengde (Zipf, 1965 [1935], s. 25-28). Det vil si at frekvente ord har en tendens til å være korte. Dette ser vi tydelig av toppen av en frekvensliste over ordformer fra elevtekstkorpuset, der bare én ordform (\ikke\) er på over 3 bokstaver.

Ordform	Antall
det	2388
er	2122
og	1674
at	1311
å	1311
som	1175
de	1158
ikke	1123
på	1117
i	1099
jeg	1023
til	934
med	874
en	841
har	705
men	687
for	629
av	570
så	564
kan	544

I muntlig språk består ingen av disse ordformene av mer enn 3 fonemer, og blant de 10 første ordformene har 8 bare 1 eller 2 fonemer. I sammenhengende tale vil dessuten flere av ordformene gjerne få en redusert form; for eksempel uttales gjerne \ikke\ bare som /ke/ med 2 fonemer. I det følgende bruker jeg imidlertid bokstaver som enhet.

Pearsons korrelasjonsanalyse for ordformtypene i hele korpuset gir $R \approx -0{,}16$ mellom ordformfrekvens og ordformlengde, noe som ikke representerer en særlig sterk sammenheng. Sammenhengen mellom ordformlengde og frekvens er imidlertid tydeligst om

man begrenser utvalget til de mest frekvente ordformene. Figur 9-1 viser korrelasjonskoeffisienten som en funksjon av hvor mange av de mest frekvente ordformene som regnes med. Grafen er tegnet bare fra og med de 15 mest frekvente ordformene; korrelasjonene for færre ordformer har store svingninger som forstyrrer mønsteret i variasjonen og gjør diagrammet vanskeligere å lese.



Figur 9-1: Korrelasjonskoeffisient for korrelasjonen mellom ordformlengde og ordformfrekvens for de 1000 mest frekvente ordformene. Diagrammet viser verdiene for Pearsons R etter hvert som analysen tar hensyn til flere og flere ordformtyper lengre nedover frekvenslista.

Figur 9-1 viser at korrelasjonen er sterkest når bare omtrent de 35 mest frekvente ordformtypene regnes med, der R nærmer seg –0,5. Deretter blir korrelasjonen raskt svakere frem til lista er ca. 150 ord lang. For lengre lister stiger den negative R svakt oppover fra ca. –0,30 og når altså –0,16 for elevtekstkorpuset som helhet. Dette virker umiddelbart ikke som en særlig sterk sammenheng. At den er signifikant, er hevet over enhver tvil; dette er jo et mønster som skriver seg fra flere tusen observasjoner. Men om den kan sies å være substansiell, er et annet spørsmål. Det er dog viktig å merke seg at sammenhengen vil virke vesentlig sterkere når man leser løpeordene i en tekst, enn i en frekvensordliste; regnet på grunnlag av løpeordene er korrelasjonen vesentlig sterkere, $R \approx -0{,}52$.

Sammenhengen mellom ordformlengde og ordformfrekvens medfører altså at det vil være en sammenheng mellom gjennomsnittlig ordlengde i en tekst og tettheten av ord som generelt er frekvente i språket. Høyere gjennomsnittlig ordlengde kan dermed gjenspeile tre eller kanskje fire litt ulike tekstlige egenskaper, som alle kan henge sammen med planlagt språkbruk:

- Lavere andel funksjonsord
- Mer spesifikke leksikalske ord
- Mer integrert ordbruk
- Mer variasjon i leksikalske ord

Funksjonsord er mer frekvente og har en tendens til å være kortere enn leksikalske ord (se diskusjonen om definisjon av leksikalske ord i 9.2.1 nedenfor). Totalt i korpuset er gjennomsnittlig ordformlengde for leksikalske ord 5,67, mens den for grammatiske ord er 2,79. Det vil si at kortere gjennomsnittlig ordformlengde i en tekst kan tyde på lavere andel av leksikalske ord i teksten, altså lavere leksikalsk tetthet, som ifølge Halliday (1989, s. 61) er et kjennetegn på en muntlig eller spontan stil. Fra frekvenslista over ser vi at alle de 20 mest frekvente ordformene har funksjonsordegenskaper[12].

Mer spesifikke leksikalske ord har en tendens til å være lengre enn mer generelle leksikalske ord. Høyere gjennomsnittlig ordlengde kan derfor tyde på et mer presist innhold. Men det kan også skrive seg fra en mer integrert ordbruk, uten at *presisjonen* nødvendigvis er høyere. Et typisk eksempel på at integrert ordbruk ikke nødvendigvis fører til høyere presisjon, er bruk av nominaliseringer eller sammensetninger til erstatning for lengre fraser; i forrige avsnitt bruker jeg for eksempel det lange ordet \funksjonsordegenskaper\, som kan representere et eksempel på integrert ordbruk, men som neppe er mer *presist* enn \egenskaper knyttet til funksjonsord\, som har lavere gjennomsnittlig ordlengde. Både høyere presisjon og integrert ordbruk blir imidlertid regnet som tegn på skriftlig, planlagt språkbruk (Biber, 1988, s. 104-; 1995, s. 141-; Chafe, 1982, s. 39-; Halliday & Matthiessen, 2004, s. 656-)

Mer variasjon i bruken av leksikalske ord vil normalt dessuten øke andelen av mindre frekvente ord, så økt variasjon kan derfor også indirekte påvirke den gjennomsnittlige ordlengden i teksten positivt.

9.1.1 Korpussøk

Analysen av ordformlengde benytter taggerprogrammet CG3 (se 6.4.2 og 8.4.1) sin definisjon av ord. Denne faller i stor grad sammen med det grafiske ordet, der et ord er en sekvens av bokstaver, tall og bindestreker adskilt av mellomrom eller annen tegnsetting. I elevtekstkorpuset finner taggerprogrammet dessuten 197 forekomster av flerordsleksemer, fordelt på 48 ulike typer. Dette er hovedsakelig leksemer av typen \for eksempel\, \i stedet for\, \til og med\, men også enkelte proprier som er skrevet i anførselstegn, som \"World of Warcraft"\ og \"Financial Times"\. Tabell 9-1 under viser alle flerordsleksemer i korpuset, samt antall forekomster i korpuset.

Tabell 9-1: Flerordsleksemer i korpuset, slik de er definert og gjenkjent av CG3-taggeren

antall	lemmaform	antall	lemmaform	antall	lemmaform
36	for eksempel	2	for resten	1	jo menn
21	i hvert fall	2	først og fremst	1	ned fra
15	blant annet	2	i så fall	1	over styr

[12] \Har\ klassifiserer jeg som enten leksikalsk eller grammatisk, avhengig av funksjonen ordet har i klaususen. Se definisjonen i 9.2.1.

antall	lemmaform	antall	lemmaform	antall	lemmaform
15	rett og slett	2	så vidt	1	på forhånd
13	stort sett	1	all verden	1	på tide
11	i stedet	1	dit hen	1	så fremt
11	i stedet for	1	f.eks.	1	så vel som
11	til og med	1	for tiden	1	til felles
5	for så vidt	1	i går	1	til stede
5	i alle fall	1	i hytt og pine	1	til syvende og sist
4	etter at	1	i kveld	1	til verks
3	bortsett fra	1	i mellomtiden	1	"Financial Times"
3	i ferd med	1	i ny og ne	1	"Mein Kampf"
3	til tider	1	i rette	1	"The Sims 2"
3	ved hjelp av	1	i tilfelle	1	"World of Warcraft"
2	den dag i dag	1	i vei	1	"mainstream operativsystem"

Oversikten viser at det er flere problemer med denne lemmatiseringen. For det første er det ikke i overensstemmelse med formålet med denne delstudien om gjennomsnittlig ordformlengde å klassifisere flere av disse sekvensene som egne leksemer, for eksempel å regne \til syvende og sist\ som en 19-bokstavers ordform.

For det andre virker det heller ikke helt valid å klassifisere \for eksempel\ som en ordform på 12 bokstaver, mens \feks\, \f.eks\ og \f. eks\ regnes som ordformer med henholdsvis 4, 5 og 6 bokstaver. Antagelig er lengden av *forkortelsene* i dette tilfellet et mer valid mål på spesifisiteten eller frekvensen av leksemet enn lengden av den uforkortede ordformen. Tabell 9-2 under viser en oversikt over forkortelser i korpuset. I tillegg til disse finnes det noen tilfeller av akronymer, som \FM\, \FN\ og \IT\, og ett spillnavn som analyseres som forkortelse av taggerprogrammet, \CS\.

Tabell 9-2: Forkortelser som taggerprogrammet har gjenkjent og tagget som forkortelser.

antall	ordform
8	ca
7	etc, etc.
1	evt.
10	feks, f.eks, f. eks
1	m.a.o.
1	m.m
1	o.l.
25	osv, osv.
8	pga, pga.

For det tredje er det flere feilparsinger i lista. De to tilfellene av \for resten\ er ikke adverber i teksten, men en del av sekvensen \for resten av livet\, og \jo menn\ er en forekomst av modalpartikkel + substantivfrase.

For det fjerde tolker taggerprogrammet flerords proprier ulikt avhengig av om de er skrevet i anførselstegn eller ikke; \World of Warcraft\ er tolket som 3 ord, mens \"World of Warcraft"\ altså er tolket som 1 ord. \"Mainstream operativsystem"\ er selvfølgelig ikke et proprium i det hele tatt, men 2 ordformer der det første er et ganske moderne importord. På

den annen side er slike tilfeller altså temmelig sjeldne, og de spiller liten praktisk rolle for denne delen av analysen. Jeg har derfor valgt å ikke løse opp disse få tilfellene i enkeltord, men forholde meg til analysen fra taggerprogrammet. Å følge taggerprogrammet slavisk gjør det også enklere å sammenligne resultater fra ulike studier som bruker det samme taggerprogrammet.

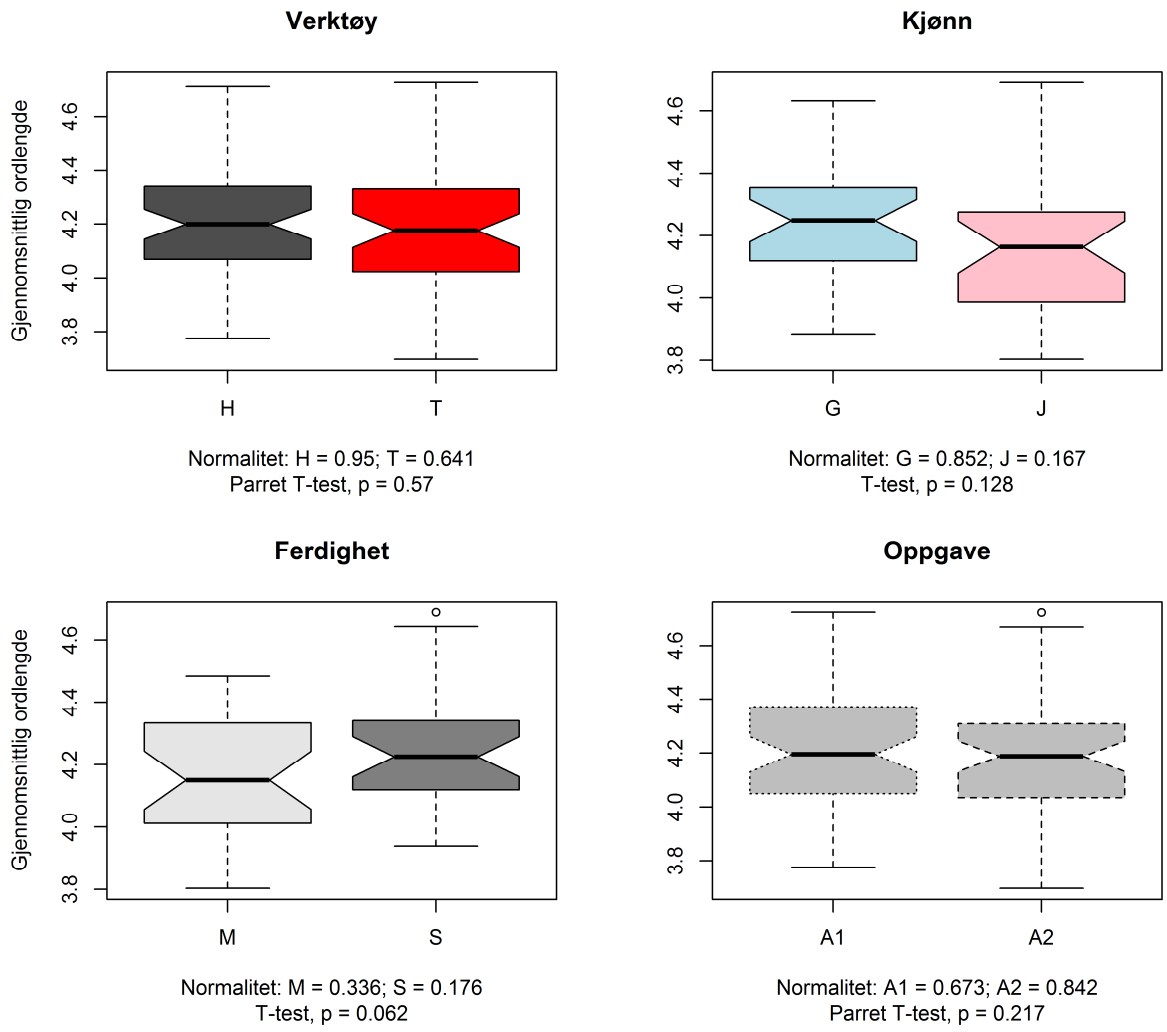
Til slutt må jeg nevne at det er 397 ordformer i korpuset der det er brukt tall-symboler i stedet for bokstaver. De vanligste ordformene er \18\, \14-15-åringene\ og \2006\, alle knyttet enten til tematikken i "Ungdomsfylla"-oppgaven eller til dateringen av sitatet i den samme oppgaven. I disse tilfellene er ordformlengden i antall grafemer kortere enn det semantiske innholdet eller den morfologiske strukturen skulle tilsi, men som et mål for den fysiske og kanskje kognitive belastningen ved å *skrive* et visst antall grafemer, er det selvfølgelig presist.

9.1.2 Deskriptiv analyse

Tabell 9-3 under viser nøkkeltallene for gjennomsnittlig ordformlengde målt i antall grafemer; siden hovedvekten av grafemene er bokstaver, har jeg valgt å bruke betegnelsen bokstaver i omtalen. Middelverdiene ligger i overkant av 4 bokstaver, med et standardavvik på noe over 0,2. Utvalget er normalfordelt, ifølge Shapiro-Wilks normalitetstest, $W \approx 0{,}990$, $p \approx 0{,}53$. Alle relevante delutvalg er også normalfordelt, ifølge samme test.

Tabell 9-3: Nøkkeltall for gjennomsnittlig ordformlengde i antall bokstaver

	middelverdi	median	sd	min	maks
Total	**4,20**	**4,19**	**0,22**	**3,69**	**4,72**
Hånd	4,20	4,19	0,20	3,77	4,71
Tast	4,19	4,18	0,24	3,69	4,72
Middels	4,15	4,14	0,20	3,75	4,61
Sterk	4,24	4,23	0,22	3,69	4,72
Gutt	4,23	4,23	0,20	3,77	4,72
Jente	4,16	4,15	0,23	3,69	4,72



Figur 9-2: Gjennomsnittlig ordformlengde. Utvalget og alle segmenter er normalfordelte, ifølge Shapiro-Wilks normalitetstest.

Figur 9-2 viser at skriveverktøy og oppgave ikke påvirker gjennomsnittlig ordlengde for korpuset som helhet. Boksdiagrammene viser at de sterke elevene og guttene i utvalget bruker litt lengre ordformer. Det er heller ingen sammenhenger mellom ordlengde og tekstlengde, R ≈ –0,06 med logaritmetransformert tekstlengde.

9.1.3 Variansanalyse

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer. Prediktorinteraksjonen er begrenset til 2 nivåer. (Se forklaringen av analysemetoden i 7.3 ovenfor.)

(83) `lm(lexD$ordlengde~(kjønn+ferdighet+lengde+forskjell)^2)`

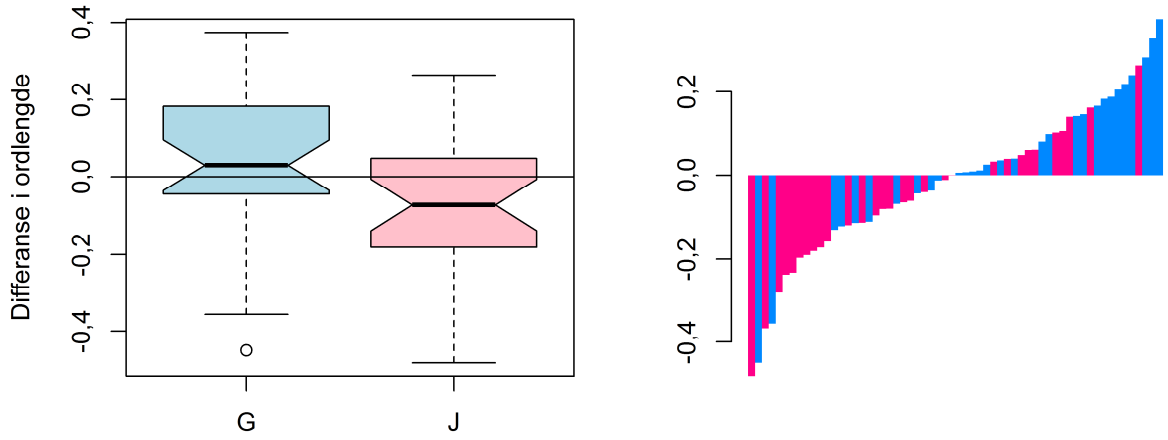
Reduksjon av den maksimale modellen gir den minimale adekvate modellen i (84):

(84) `lm(formula = lexD$ordlengde ~ kjønn)`

```
             Df Sum Sq Mean Sq F value Pr(>F)
kjønn         1 0.2037  0.2037   6.789 0.0116 *
Residuals    58 1.7400  0.0300
---
Multiple R-squared:  0.1048,    Adjusted R-squared:  0.08936
F-statistic: 6.789 on 1 and 58 DF,  p-value: 0.01163
```

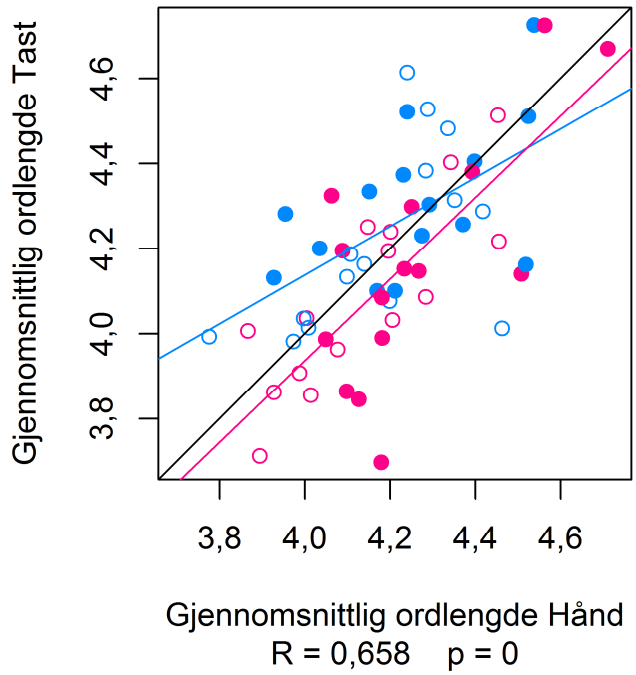
`gvlma` (se 7.2.2.4) viser at premissene for anova er oppfylt, men at skjevhetsfaktoren er på grensen av det som regnes som akseptabelt. (Se appendiks A4.)

Kjønn er altså den eneste signifikante faktoren i modellen, og figur 9-3 viser at jenter har en tendens til lavere gjennomsnittlig ordlengde i tastetekster, mens gutter har en tendens til høyere. (Se også figur 10-58 på side 239, som viser et boksdiagram over både hånd- og tastetekster for gutter og jenter.) Skreddiagrammet til høyre i figuren viser at guttene dominerer sterkt blant dem som har mye høyere ordlengde i tastetekstene, mens 2 gutter med lave verdier forstyrrer jentedominansen noe i den andre enden av skalaen.



Figur 9-3: Differanse i ordformlengde fordelt på kjønn. Til venstre et boksdiagram. Til høyre et skreddiagram, der verdiene er sortert etter størrelse.

Det som imidlertid skiller denne variabelen fra mange andre av variablene i denne avhandlingen, er den ganske sterke korrelasjonen mellom håndtekstene og tastetekstene, $R \approx 0{,}66$, som figur 9-4 viser (CI(95%) ≈ [0,48, 0,78]). For jentene er $R \approx 0{,}76$ (CI (95%) ≈ [0,56, 0,88]) og for guttene $R \approx 0{,}56$ (CI (95%) ≈ [0,25, 0,77]). Verdien for jentene er svært høy og indikerer at jentene kanskje i enda større grad enn guttene har funnet sin stil med hensyn til denne variabelen, selv om den altså blir påvirket noe av verktøy. Dessuten er det ganske stor overlapp mellom konfidensintervallene, så det er usikkert om forskjellen mellom kjønnene er reell.



Figur 9-4: Sammenheng mellom gjennomsnittlig ordformlengde i håndtekster og tastetekster for gutter og jenter, middels og sterke. Middels elever er representert med åpne sirkler, sterke elever med fylte sirkler. Gutter med lyseblått og jenter med rosa.

Figur 9-4 viser også at de 8 tastetekstene med lavest gjennomsnittlig ordlengde er skrevet av jenter, 5 middels og 3 sterke. Totalt er 10 av de 11 tekstene med lavest ordlengde skrevet av jenter; to jenter (elev-id 248 og 249), begge med karakteren 4, har begge sine tekster i dette sjiktet. Disse 10 jentetekstene utgjør mesteparten av forklaringen på kjønnsforskjellen i gjennomsnittlig ordlengde.

Figur 9-4 viser også en annen kjønnsforskjell, som ikke går fram av de andre diagrammene. Nesten alle de jentene som har vesentlig lavere gjennomsnittlig ordlengde i tastetekstene, altså stor negativ avstand fra korrelasjonslinjen, har relativt lave verdier også i håndtekstene. Men de 2 guttene med størst negativ differanse har begge blant de høyeste håndtekstverdiene, og de har heller ikke særlig lave verdier i tastetekstene. Denne tendensen understreker at lav gjennomsnittlig ordlengde synes å være et jente-trekk.

9.1.4 Oppsummering og diskusjon

Analysene viser at jentene tenderer mot en mer spontan stil i tastetekstene, mens guttene tenderer mot en mer planlagt stil i tastetekstene. Tendensen er sterkest hos jentene, kanskje først og fremst på grunn av en gruppe jenter som har spesielt mye lavere verdier i tastetekstene. Dette er en tendens vi kommer til å se i flere av variablene i de kommende kapitlene, både leksikalske og syntaktiske.

Ordlengde korrelerer ganske sterkt med både leksikalsk tetthet og variablene som er relatert til leksikalsk variasjon, slik det blant annet går fram av korrelasjonskoeffisienter som er

gjengitt i 10.6. Men ordlengde kan også potensielt være relatert til tekstegenskaper som ikke henger sammen med leksikalsk tetthet og TTR, nemlig spesifisitet og integrerthet.

På grunn av den åpenbare sammenhengene med leksikalsk tetthet, bekreftet av sterk korrelasjon ($R \approx 0{,}68$), kan det tenkes at kjønnseffekten vi har funnet, først og fremst er en effekt knyttet til leksikalsk tetthet. Det vil derfor være interessant å gjøre tilsvarende analyse av ordlengde der bare de leksikalske ordene inngår i beregningene, for å se om ordlengden også representerer andre typer egenskaper. Dette kommer jeg tilbake til i 9.3.2, i en diskusjon knyttet til leksikalsk spesifisitet.

9.2 Leksikalsk tetthet

9.2.1 Definisjon

Leksikalsk tetthet er knyttet til frekvens av leksikalske løpeord. Michael Halliday fremhever at klaususstrukturen i skriftlig språk typisk er enklere enn i muntlig språk, men at innholdet er "tettere pakket" i leksikalske enheter (Halliday, 1987, 1989; Halliday & Matthiessen, 2004). På bakgrunn av dette karakteriserer han muntlig språk som dynamisk og skriftlig språk som statisk og sier at skriftlig, planlagt språk er preget av høyere leksikalsk tetthet enn muntlig, spontant språk. Leksikalsk tetthet regner han ut som gjennomsnittlig antall leksikalske ord per klausus[13] (1979, s. 49; 1989, s. 64; 2004, s. 655), men vanligere er det nok å regne leksikalsk tetthet som andelen leksikalske ord av tekstens løpeord (Johansson, 2008, s. 67; Malvern, Richards, Chipere, & Durán, 2004, s. 3; O'Loughlin, 1995; Stubbs, 1996, s. 72), slik Halliday også gjør ved ett tilfelle (1987, s. 60). De to målestokkene representerer to konseptuelt ganske ulike variabler; antall leksikalske ord per klausus sier noe om hvor mye innhold som pakkes i hver klausus, noe som i praksis oftest svarer til hvert hovedverb, mens antall leksikalske ord per løpeord sier noe om hvor tett innholdet pakkes, eller i hvor stor grad funksjonsord brukes til å få fram forholdet mellom de leksikalske ordene.[14]

Når det gjelder definisjonen av leksikalske og grammatiske enheter, blir den av Halliday og visse andre forfattere først og fremst knyttet til at leksikalske enheter er medlemmer av åpne klasser, mens grammatiske enheter er medlemmer av lukkede klasser (Borgstrøm, 1973, s. 61-75; Halliday, 1989, s. 63; Stubbs, 1996, s. 72). Dette er imidlertid en definisjon som er avhengig av hvordan man definerer og avgrenser de enkelte ordklasser, og andre forsøk på definisjoner er tettere knyttet til ordenes semantiske og formelle egenskaper. Stubbs (s. 71)

---

[13] Hallidays klaususbegrep for engelsk er noe forskjellig fra mitt, så konkrete verdier vil ikke være direkte sammenlignbare. I praksis er imidlertid ikke forskjellen så stor.

[14] Leksikalsk tetthet svarer således delvis, men bare delvis, til syntetiske virkemidler i motsetning til analytiske. I \Knut slo Arne\ er leksikalsk tetthet 1, men mye av informasjonen er pakket analytisk, bare ikke med funksjonsord.

peker på at leksikalske ord uttrykker innhold, mens grammatiske ord forbinder leksikalske ord med hverandre. Kulbrandstad (2005, s. 110) sier også at innholdsord "har et eget, selvstendig betydningsinnhold", mens funksjonsord "angir [...] forholdet mellom innholdsorda i setningen og viser dermed hvordan betydningsinnholdet i disse orda føyes sammen til en helhet." Alle disse kildene peker imidlertid på at det ikke går an å trekke noe absolutt skille mellom leksikalske og grammatiske ord, men at det snarere er snakk om et kontinuum (Halliday, 1989, s. 63), et spørsmål om grader (L. A. Kulbrandstad, 2005, s. 111), noen som står i en mellomstilling (Borgstrøm, 1973, s. 63), eller *rough categories* (Stubbs, 1996, s. 71).

Den enkleste tilnærmingen til avgrensningen av leksikalske ord ville være å følge tradisjonell ordklasseinndeling slik den er operasjonalisert i taggerprogrammet. I så fall kunne substantiver, verb, adjektiver (Faarlund, et al., 1997, s. 21)[15], adverb og eventuelt interjeksjoner (L. A. Kulbrandstad, 2005, s. 111) regnes som åpne klasser og dermed alle deres medlemmer for leksikalske ord. Dette er imidlertid en utilfredsstillende løsning fordi det finnes medlemmer av flere av disse ordklassene som har fremtredende grammatiske egenskaper. Siden nettopp disse ordene er relativt frekvente, vil det i en kvantitativ undersøkelse som denne ha temmelig store konsekvenser å regne dem blant de leksikalske ordene.

Nettopp frekvens i bruk kan være en pekepinn på om et leksem fra en åpen ordklasse bør regnes som grammatisk. Blant substantivene er det ingen av de mest frekvente lemmaene som synes å ha fått en dominerende grammatisk funksjon. Blant verbene er det derimot flere interessante eksempler. Tabell 9-4 under viser alle verblemma som forekommer minst 100 ganger i korpuset:

Tabell 9-4: De mest frekvente verblemmaene i korpuset

antall	lemma
2645	være
953	ha
583	kunne
562	lese
453	bli
395	gjøre
382	drikke
365	få
305	ville
288	si
265	komme
236	skulle

[15] Faarlund, et al. (1997, s. 24) holder dessuten preposisjoner for å være leksikalske ord, men ser ut til å være alene om dette. Næs opererer ikke med skillet mellom leksikalske og grammatiske ord, men mellom åpne og lukkede klasser.

antall	lemma
234	drive
230	tro
217	gjø
171	burde
171	sitte
170	bruke
159	se
137	mene
131	kjøpe
131	måtte
129	ta
118	vite
117	gå
106	synes
101	like
100	gi

Blant disse ser vi at \være\ er desidert mest frekvent, mens begge hjelpeverbene \bli\ (for passiv) og \ha\ (for perfektum) også kommer høyt opp på lista. Dessuten er alle de fem tradisjonelle modalverbene \kunne\, \ville\, \burde\, \skulle\, \måtte\ representert, i tillegg til \få\, som i mange konstruksjoner fungerer som et modalt hjelpeverb. Dessuten ser vi noen leksikalske verb som er knyttet til oppgavens tematikk, \lese\, \drikke\, \drive\ og \sitte\, og som formodentlig ikke er like frekvente i andre korpus. Lemmaformen \gjø\ er en feil-lemmatisering av \gjør\, og leksemet er i realiteten ikke til stede i korpuset i det hele tatt.[16] Utover disse finner vi noen interessante verb med spesielle egenskaper.

I engelsk er *DO*\ et typisk grammatisk ord, men i norsk har ikke \GJØRE\ den samme funksjonen. Dog brukes \GJØRE\ også i norsk som et pro-ord på linje med pronomener som \den\ og proadverb som \SLIK\, for eksempel i (85) der \gjorde\ representerer \DRIKKE\. Men \GJØRE\ kan også ha en mer selvstendig, skjønt vid, leksikalsk betydning, som i (86)

(85) Vi drikker mer alkohol enn vi gjorde tidligere. [A2-264]
(86) Jeg vet at mange er nok nysgjerrig på hva alkohol gjør med kroppen deres, [A2-305]

\KOMME\ kan ha temmelig grammatiske funksjoner, som for eksempel som fremtidsmarkør sammen med partikkelen \til\ (87), eller som i (88), der \kommer\ inngår i en klausus som nærmest fungerer som en temamarkør. I mange tilfeller i elevtekstkorpuset brukes \KOMME\ som del av en markør for direkte eller indirekte tale, ofte som i (89) i kombinasjon med \med\, men ikke alltid (90). I dette tilfellet har \KOMME\ nærmest et preg av å være første ledd i en pseudo-koordinasjon.

[16] Taggerprogrammet har i disse tilfellene lemmatisert \gjør\ både som \gjøre\ og som \gjø\. Tallene for \gjøre\ er dermed riktige, mens tallene for \gjø\ kan ses bort fra.

(87) Da vet man ikke hvordan man kommer til å reagere på alkohol. [A2-313]
(88) Øl og rusbrus er en ting, men når det kommer til sprit, kan ting fort gå litt over styr når de er på denne alderen. [A2-297]
(89) Nå skriver jeg personlig til deg som kom med denne påstanden; [A1-301]
(90) For ikke så lenge siden kom lillesøsteren hennes som er femten, og lurer på om venninnen min kunne kjøpe henne alkohol innimellom. [A2-235]

\DRIVE\ er et verb som kan brukes i pseudokoordinasjon, som i (91), men dette er faktisk det eneste tilfellet i korpuset. Nesten alle de andre tilfellene er i sammenheng med \med\ eller \på\ og \data\, \PC\, \dataspill\ eller lignende, og jeg regner med at de fleste av disse formuleringene er sterkt inspirert av oppgaveformuleringen.

(91) For det er ikke mange ti-elleve-åringer som driver og smugdrikker. [A2-292]

Når det gjelder leksemet \SITTE\, finnes det 31 tilfeller av pseudokoordinasjon, ofte i sammenheng med \spille\ eller \lese\, men de utgjør altså ikke noen stor andel av de totalt 171 forekomstene av \sitte\, under 20 %.

Som Halliday (1989, s. 63) sier, har det ikke nødvendigvis så stor betydningen akkurat hvor man trekker skillelinjen mellom leksikalske og grammatiske ord, så lenge det gjøres konsekvent, og jeg har valgt å inkludere \være\ og de 5 tradisjonelle modalverbene blant de grammatiske ordene. Derimot har jeg valgt å regne \gjøre\ og \komme\ som leksikalske, selv om de har temmelig frekvente forekomster med lite leksikalsk innhold. Når det gjelder verbene \ha\, \få\ og \bli\, som har både leksikalske og grammatiske varianter (Borgstrøm, 1973, s. 63-65), regner jeg dem som grammatiske der de fungerer som hjelpeverb sammen med infinitte verbformer, og som leksikalske ellers.

Blant de mest frekvente adjektivene er det \mye\ og \mange\ som skiller seg ut som særlig frekvente. Samtlige forekomster av \mye\ og \mange\ eller bøyningsformene \mer\, \mest\ og \meste\ og \flere\, \flest\ og \fleste\ er tagget som adjektiv.

Tabell 9-5: De mest frekvente adjektivene i korpuset

antall	lemmaform
670	mye
533	mange
169	hel
153	god
153	litt
118	stor
112	gammel
107	ung
105	enkel
101	enkelt
101	liten
98	gammal
92	veldig
86	ofte

antall	lemmaform
85	negativ
76	full

Dette er i tråd med en morfologisk definisjon av adjektiv, der gradbøyning er et sentralt kriterium. *Norsk Referansegrammatikk* (Faarlund, et al., 1997, s. 354) kategoriserer da også disse to leksemene som adjektiver. Kulbrandstad (s. 173) nevner \MYE\ blant adverbene, men \MANGE\ blant adjektivene, mens Næs (s. 216-217) lister både \MYE\ og \MANGE\ blant de adjektiviske pronomen og sier at \MANGE\ (blant annet) er flertall til \MYE\.

Syntaktisk (og semantisk) er imidlertid disse ordene problematiske, fordi de har flere ulike funksjoner, (92) – (95). Klarest grammatisk er de 111 tilfellene der \mer\ og \mest\ er brukt til gradbøyning av adjektiv. Men også \mye\ som gradsadverb (Næs, 1965, s. 217) (92) og \mange\ og \mye\ i kvantor-funksjon (Faarlund, et al., 1997, s. 218) (93) og (94) har fremtredende grammatiske egenskaper, mens adverbial bruk av \mye\ (95) har sterkere preg av leksikalsk innhold. \Mange\ kan også ha en mer beskrivende, adjektivisk funksjon (96), men det er vanskelig å finne entydige eksempler på det i elevtekstkorpuset.

(92) Det er mye artigere å sitte på en data enn å lese masse bokstaver. [A1-202]
(93) Jeg vet om mange gutter som liker å lese bøker, [A1-212]
(94) Ungdom får mye kunnskap ut av internett, [A1-222]
(95) men guttene spiller mye på data, [A1-205]
(96) Han snakket om sine mange problemer. [BUJ]

Generelt er det her snakk om glidende overganger, og både økonomisk og prinsipielt ville det være vanskelig å analysere hver forekomst og skille dem dikotomt i leksikalske og grammatiske. Derimot er det faktum at \MYE\ og \MANGE\ er 3-4 ganger så frekvente som det tredje adjektivet på lista, et sterkt indisium på at disse leksemene har fremtredende grammatiske egenskaper. Jeg bruker derfor forskjellen i frekvens som hovedargument for å skille ut alle forekomster av leksemene \MYE\ og \MANGE\ som grammatiske ord. Siden de er såpass frekvente, er dette et valg som kan ha en del innvirkning på de statistiske analysene. Også \hele\ kan ha preg av å være en kvantor, men for \hel\ har jeg valgt å bruke frekvensargumentet til å regne det blant de leksikalske ordene, selv om det altså er nummer 3 på lista over frekvente adjektivlemma, eller øverst når \mye\ og \mange\ ikke regnes som adjektiv.

Lista av adjektiv illustrerer også at en del leksikalske ord som tradisjonelt er regnet som adverb, av taggeren blir klassifisert som adjektiver, jf. \litt\, \veldig\ og \ofte\. De mange tilfellene av \enkelt\ eller \enkel\ skyldes at \enkelt\ er brukt i den ene oppgaveteksten.

Halliday (1989, s. 63) og Kulbrandstad (2005, s. 111) peker spesielt på at adverb er en ordklasse som inneholder både leksikalske og grammatiske ord, og en oversikt over de mest frekvente adverbene viser at den inneholder mange ord med klare grammatiske egenskaper og få ord med fremtredende leksikalske egenskaper. Det er 154 adverbtyper og 4664 eksemplarer i korpuset, hvorav omtrent en fjerdedel er \ikke\.

Tabell 9-6: De mest frekvente adverbene i korpuset

antall	lemma
1124	ikke
543	så
261	også
217	bare
195	da
194	jo
176	kanskje
162	nok
117	hvor
105	slik
92	like
86	heller
81	nå
77	selv
75	vel
51	aldri
49	for
49	sånn
47	alltid
44	derfor
44	hvorfor
43	hvordan
41	ganske
36	for eksempel
30	siden

Det finnes mange leksikalske adverb i korpuset, som \tydelig\, \ulovlig\ og \umulig\ med 5 forekomster hver, men klassen av adverber er temmelig sammensatt og illustrerer tydelig hvordan "leksikalske ord" er en prototypisk klasse med glidende overganger. De fleste av de mest frekvente adverbene er ganske typisk grammatiske ord, men også langt ned på lista er det en stor andel av ord med klare grammatiske trekk. Tabell 9-6 ovenfor viser de adverbene som forekommer minst 30 ganger i korpuset.

For å unngå individuell behandling av denne store klassen av ord, og med den viktige begrunnelsen at de fleste av de mest frekvente ordene i klassen er grammatiske, har jeg valgt å utelate alle adverbene fra klassen av leksikalske ord.

Kulbrandstad (2005, s. 110) nevner også interjeksjonene blant de åpne ordklassene, men de refererer ikke til "noe i verden", i hvert fall ikke funksjonelt, og jeg regner dem blant de grammatiske.

Tabell 9-7: Alle ord som er tagget som interjeksjoner i korpuset

antall	lemmaform
25	ja
18	nei

antall	lemmaform
7	jo
3	nå
2	å
1	"guttegreie"
1	gøyest
1	hei
1	herregud
1	kommunikasjonsverktøyer
1	neida
1	vips

I elevtekstkorpuset dreier det seg i all hovedsak om \ja\, \nei\ og \jo\, og man kan dermed bruke argumentet om at interjeksjonene hovedsakelig er en lukket klasse i dette korpuset, for å kategorisere dem som grammatiske ord. Tabell 9-7 illustrerer at det også når det gjelder interjeksjoner, er noe unøyaktighet i taggingen.

Halliday (1989, s. 63) legger vekt på å segmentere i "*lexical items*" og ikke i (grafiske) ord, og han trekker spesielt frem verbpartikler, som han vil analysere som del av den leksikalske enheten med verbet som kjernen. I likhet med skillelinjen mellom leksikalsk og grammatisk er det både prinsipielt og praktisk vanskelig å segmentere potensielt ikke-kontinuerlige flerordsleksemer, og jeg har valgt å holde meg til taggerprogrammets segmentering i denne undersøkelsen.

9.2.2 Korpussøk

Som vist i 9.2.1 er det ingen entydig, avgrensende definisjon av leksikalske ord. (97) – (102) viser korpussøkene som er aktuelle.

(97) `[features=("subst") & word!="i"%c] | [features=("verb") & lemma!="være"] | [features=("adj") & word!="av"%c & lemma!="mye" & lemma!="mange"]`

(98) `([lemma="ha"]|[lemma="bli"])[!<clause> & !</clause> & !<t-unit>]*[features=("perf-part")]`

(99) `[lemma="kunne"]|[lemma="skulle"]|[lemma="ville"]|[lemma="måtte"]|[lemma="burde"]`

(100) `[features=("<aux1/infinitiv>")]`

(101) `[features=("verb") & lemma="få"][lemma!="å"]{0,1}{[features=("@iv") & lemma!="drikke"] (47 treff)`

(102) `[features=("adv")]`

(97) trekker ut alle ordene fra de tre viktigste leksikalske ordklassene. Det er viktig å ikke bruke ett søk for hver ordklasse og deretter summere tallene, for taggerprogrammets garderte lemmatisering (se 6.4.2) ville i så fall resultere i at mange ord telles flere ganger om man

ikke fjernet dublettene.[17] Blant substantiver og adjektiver har jeg fjernet to frekvente feiltagginger automatisk (\i\ og \av\), mens jeg har fjernet alle forekomstene av \være\ fra verbene og alle forekomstene av \mye\ og \mange\ fra adjektivene.

(98) trekker ut lemmaene \ha\ og \bli\ som hjelpeverb. Dette er gjort gjennom å finne forekomster som står etterfulgt av perfektum partisipp i samme klausus. (Se 9.2.1.)

(99) og (100) er to alternative og likeverdige måter å trekke ut modale hjelpeverb på.

\Få\ er ikke tagget som modalt hjelpeverb i korpuset, men har i visse kontekster egenskaper som åpenbart er nært beslektet med hjelpeverb, både formelt og semantisk. Søket i (101), som finner alle forekomster av verb-leksemet \få\ etterfulgt av et infinitt verbal, med maksimalt ett annet ord som ikke er \å\, imellom, får 47 treff. Muligheten for at det finitte verbalet er \drikke\, er fjernet fra søket, fordi 9 av 10 tilfeller av \drikke\ i denne posisjonen er substantiv selv om de er tagget som verb. Det er 8 tilfeller der subjektet kommer mellom hjelpeverbet og hovedverbet, og 3 treff der et setningsadverbial kommer imellom. Å tillate to ord imellom fører til ytterligere 3 treff, hvorav ett er et feiltreff. Tre ord imellom fører til ytterligere 3 korrekte treff og 4 feiltreff. Å sette grensen ved 1 ord ser dermed ut til å være et rimelig kompromiss som ikke medfører behov for manuell rensing.

Det er komplisert og lite hensiktsmessig å gjøre denne typen komplekse søk direkte i Corpuscle så lenge korpussystemet ikke støtter mengdesubtraksjon. Men det er også risikabelt å gjøre subtraksjonen direkte på tallvektorene fra søkeresultatene, fordi feil som skyldes gardering i taggingen, feiltagging eller feiltreff i korpus, kan forårsake forsterkede feilmønstre ved at man subtraherer tilfeller i minuenden som ikke er reelle forekomster. Det er derfor bedre å importere korpusposisjonen, som er et slags identifikasjonsmerke for hvert ord i korpuset, og bruke korpusposisjonene som grunnlag for subtraksjon. Det er dette jeg har gjort for å finne leksikalske ord; resultatene fra (98), (99) og (101) er trukket fra (97).

Leksikalske ord er dermed summen av alle substantiv, alle verb bortsett fra \være\, de modale hjelpeverbene, og \ha\, \bli\ og \få\ som hjelpeverb, og alle adjektiv bortsett fra lemmaene \mye\ og \mange\. Variabelen leksikalsk tetthet er antall leksikalske ord dividert på antall ord i teksten. (Se også 11.1.2 om klausal lengde.)

9.2.3 Deskriptiv analyse

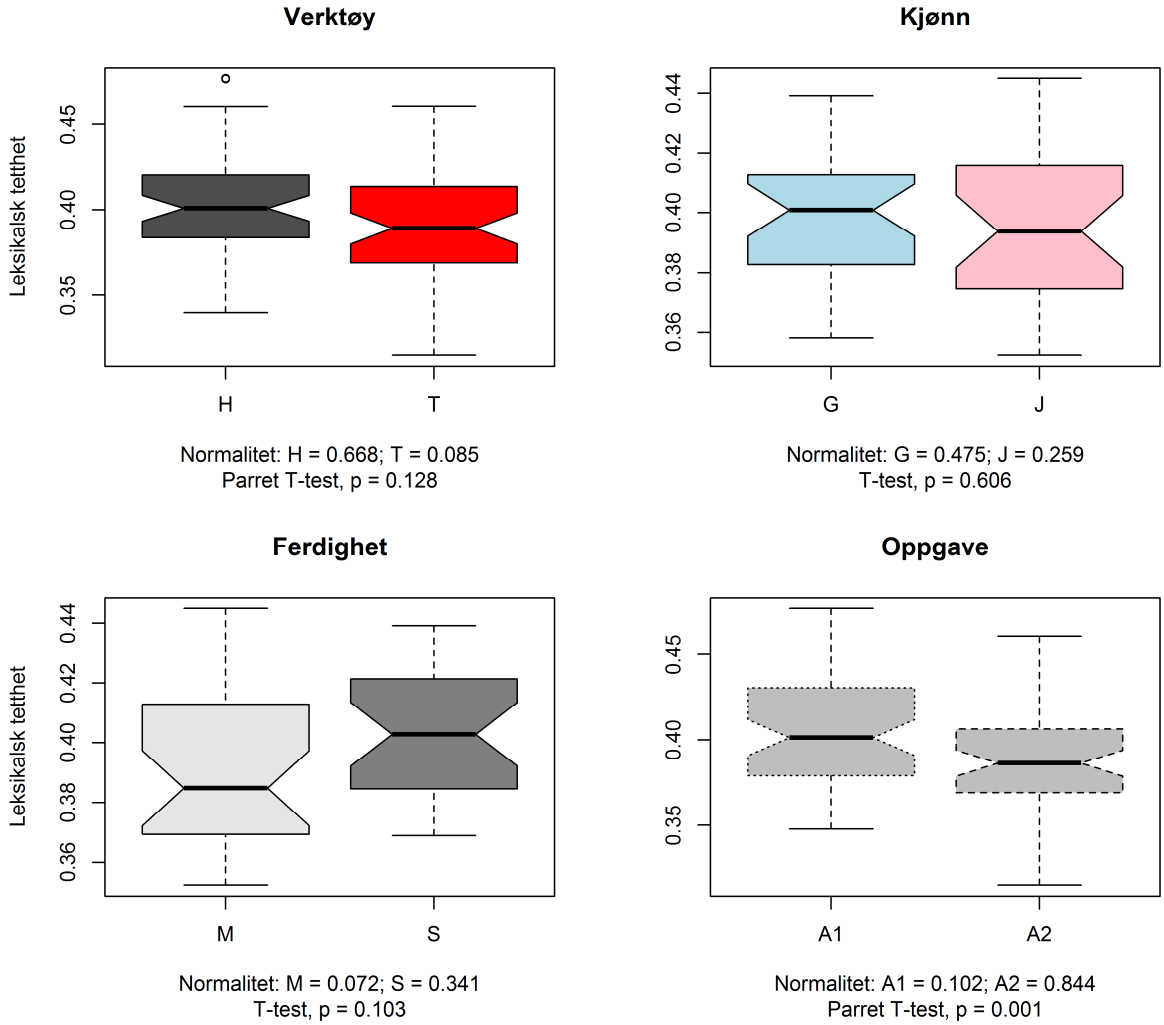
Tabell 9-8 under viser at de leksikalske ordene typisk utgjør i underkant av 40 % av løpeordene i tekstene, men det er betydelig variasjon, fra 32 % til 48 %. Utvalget og de relevante delutvalgene er normalfordelte ifølge Shapiro-Wilks normalitetstest.

[17] Se også forklaringen av metoden om å bruke unike korpusposisjoner senere i underkapitlet.

Tabell 9-8: Nøkkeltall for leksikalsk tetthet

	middel	median	sd	min	maks
Total	**0,397**	**0,394**	**0,032**	**0,315**	**0,477**
Hånd	0,401	0,401	0,031	0,340	0,477
Tast	0,393	0,389	0,033	0,315	0,461
Middels	0,392	0,384	0,036	0,315	0,477
Sterk	0,402	0,400	0,027	0,339	0,461
Gutt	0,399	0,398	0,029	0,340	0,461
Jente	0,396	0,391	0,035	0,315	0,477

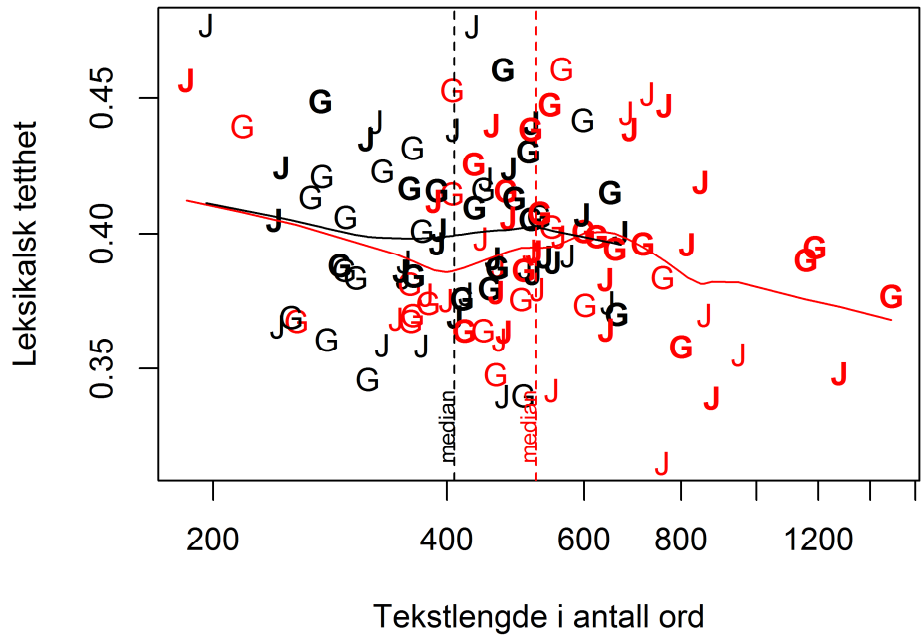
Figur 9-5 under viser at det ikke er forskjeller mellom skriveverktøy, kjønn eller skriveferdighet, men at "Bøker eller data"-tekstene har vesentlig høyere leksikalsk tetthet enn "Ungdomsfylla"-tekstene, d ≈ 0.58.



Figur 9-5: Leksikalsk tetthet fordelt etter fire faktorer

Figur 9-6 nedenfor viser at det er en viss tendens til negativ korrelasjon mellom leksikalsk tetthet og tekstlengde i tastetekstene, R ≈ –0,19, mens det ikke er noen slik effekt for håndtekstene. Det er svært svak korrelasjon mellom hånd- og tastetekster, R ≈ 0,137, noe

som tyder på enten at elevene i liten grad har noe innarbeidet skrivemønster når det gjelder denne variabelen, eller at leksikalsk tetthet i større grad enn for eksempel ordlengde (9.1.2) blir påvirket av eksterne faktorer, som emne eller elevens forhold til emnet. I den sammenhengen er det påfallende forskjell mellom oppgavens innvirkning på henholdsvis ordlengde og leksikalsk tetthet; på ordlengde har oppgaven ingen tydelig innvirkning. Oppgavens innvirkning på leksikalsk tetthet kan også bety at korrelasjonen med tekstlengde i realiteten er en effekt av oppgaven, ettersom "Ungdomsfylla"-tekstene både har større lengde og lavere leksikalsk tetthet enn "Bøker eller data"-tekstene.



Figur 9-6: Leksikalsk tetthet og sammenheng med tekstlengde. For tastetekstene er R ≈ –0,19.

9.2.4 Variansanalyse

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som respons og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer. Interaksjonen er begrenset til 2 nivåer. Analysen er altså utført på nøyaktig samme måte som for gjennomsnittlig ordlengde i 9.1.3:

(103)
```
lm(lexD$leksordF ~ (kjønn+ferdighet+lengde+forskjell)^2)
```

Resultatet er en minimal adekvat modell med svakt signifikant resultat ($F \approx 3{,}1$, $p < 0{,}05$) med interaksjonen mellom skriveferdighet og forskjell i tekstlengde som eneste signifikante faktor.

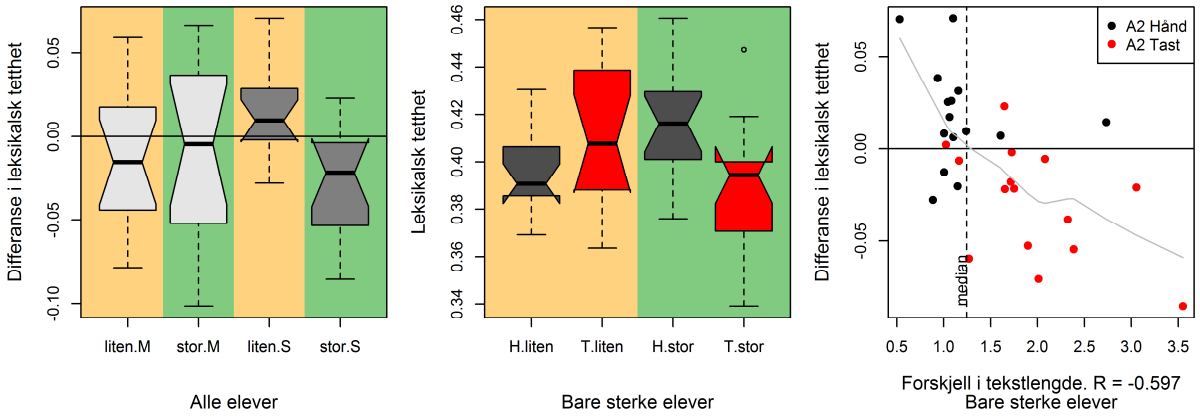
(104)
```
lm(formula = lexD$leksordF ~ ferdighet + forskjell +
ferdighet:forskjell)

                    Df  Sum Sq  Mean Sq F value Pr(>F)
ferdighet            1 0.00046 0.000464   0.292 0.5913
forskjell            1 0.00527 0.005269   3.312 0.0741 .
```

```
ferdighet:forskjell  1 0.00883 0.008830   5.550 0.0220 *
Residuals           56 0.08910 0.001591
---
Multiple R-squared:  0.1405,    Adjusted R-squared:  0.09444
F-statistic: 3.051 on 3 and 56 DF,  p-value: 0.03588
```

`Gvlma` (se 7.2.2.4) viser at premissene for anova er oppfylt. (Se appendiks A4.)

Post-hoc-testing med Tukeys HSD-test gir svakt signifikante forskjeller ($p < 0{,}05$) mellom sterke elever med stor og liten forskjell i tekstlengde. (Se appendiks A5.) Figur 9-7 nedenfor viser hva effektene består i. Diagrammet til venstre viser at det blant sterke elever er slik at de som ikke skriver vesentlig lengre på tastatur, har høyere leksikalsk tetthet i tastetekstene, mens de som skriver vesentlig lengre på tastatur, har lavere leksikalsk tetthet i tastetekstene. Diagrammet i midten demonstrerer den store ulikheten mellom de to delutvalgene av sterke elever. Ikke bare er differansene ulike, men de to gruppene har også ganske store forskjeller i typiske verdier for både håndtekstene og tastetekstene. Dette gjør det nærliggende å tenke seg at forskjellene har en viss sammenheng med oppgavetypene, ettersom vi vet at det er kraftige interaksjoner mellom tekstlengde, verktøy og oppgave (8.4.2).



Figur 9-7: Differanse i leksikalsk tetthet. Diagrammet til venstre viser differansen etter faktorene skriveferdighet og forskjell i tekstlengde. Diagrammet i midten viser leksikalsk tetthet etter faktorene skriveverktøy og forskjell i tekstlengde, men bare for sterke elever. Punktdiagrammet til høyre viser verdiene for differanse i leksikalsk tetthet som funksjon av forskjell i tekstlengde, men bare for sterke elever, og skilt etter hvilket verktøy eleven har skrevet oppgave A2 med.

Diagrammet til høyre viser imidlertid at det er en temmelig jevn tendens til fallende leksikalsk tetthet med økende forskjell i tekstlengde, noe som kunne støtte oppunder at forskjellen i tekstlengde er en faktisk faktor. Likevel ser vi her også den ekstremt skjeve fordelingen mellom de to oppgavene blant de sterke elevene; 13 av 15 elever som har skrevet mye lengre på tastatur enn for hånd, har skrevet A2-oppgaven på tastatur. Denne interaksjonen mellom tekstlengde og oppgavetekst gjør det vanskelig å konkludere om forskjellen i leksikalsk tetthet er sterkest knyttet til forskjell i tekstlengde eller til oppgavetekst. (Se også diskusjonen i 9.2.5.)

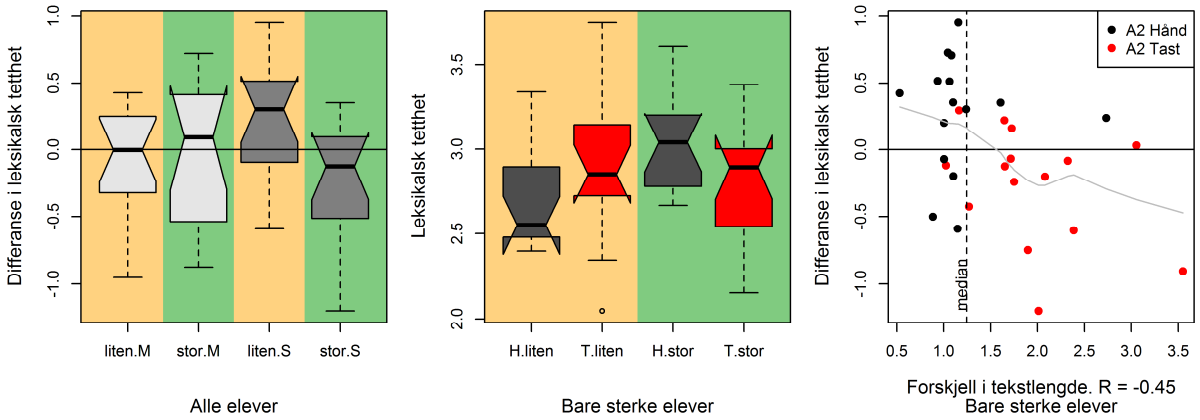
9.2.5 Oppsummering og diskusjon

Det er ingen signifikante resultater for middels elever. Sterke elever, derimot, ser ut til å dele seg i to grupper. Noen av dem benytter den økte hastigheten og/eller funksjonaliteten i tekstbehandlingsverktøyet til å skrive mer informasjonstett, mens andre benytter den økte hastigheten først og fremst til å skrive lengre, noe som resulterer i mindre informasjonstette tekster. En nærmere analyse avslører imidlertid at effekten kanskje hovedsakelig skriver seg fra forskjeller mellom de to oppgavene. Det er ingen effekter av kjønn eller totallengde for denne variabelen og heller ingen allmenn tendens til forskjell mellom verktøyene. Til tross for at korrelasjonen mellom gjennomsnittlig ordlengde og leksikalsk tetthet som ventet er temmelig sterk ($R \approx 0{,}68$, se også 10.6), er altså effektene for de to variablene tilsynelatende helt ulike. Dette tyder med andre ord på at kjønnseffekten vi fant for ordlengde (9.1.4), er knyttet til en annen type egenskap enn leksikalsk tetthet.

Leksikalsk tetthet er en variabel med flere usikkerhetsmomenter. Teorigrunnlaget er svakt for å hypotetisere rundt variabelen. For eksempel holder Biber seg til TTR og ordlengde som mål for variasjon og spesifisitet og bruker ikke leksikalsk tetthet i det hele tatt. Men dessuten er det også en rekke noe mer tekniske problemer knyttet til variabelen.

For det første er det spørsmålet om målestokk. Halliday velger å regne leksikalsk tetthet som antall leksikalske enheter per *ranking clause*. Prinsipielt er dette en ganske annen variabel enn andel leksikalske ord blant løpeordene, ettersom det måler hvor mye leksikalsk informasjon man putter inn i et syntaktisk segment, uten å ta hensyn til hvor mange ikke-leksikalske ord som også finnes i segmentet. Det innebærer altså at for eksempel mange adverbialledd med preposisjonsfraser med nominale utfyllinger vil øke Hallidays versjon av leksikalsk tetthet, selv om frasene også skulle inneholde mange funksjonsord. Og syntaktiske ledd med mange grammatiske ord vil ikke i seg selv senke den leksikalske tettheten dersom det samtidig er leksikalske ord til stede. Hallidays versjon av leksikalsk tetthet er dermed på mange måter mer en syntaktisk variabel enn en leksikalsk variabel.

Hallidays klaususbegrep er knyttet bare til verbalneksus og ikke til finitthet, og det er dermed ikke helt det samme klaususbegrepet som det jeg bruker i denne avhandlingen. Med den annoteringen som er brukt i korpuset, er det vanskelig å basere beregningene på Hallidays klaususbegrep. Jeg har imidlertid sammenlignet leksikalsk tetthet per løpeord med leksikalsk tetthet per finitt klausus i elevtekstkorpuset, og de to variablene samvarierer i ganske stor grad. Korrelasjonen mellom dem er relativt sterk, $R \approx 0{,}79$, men ikke sterk nok til å hevde at disse to variantene konseptuelt i realiteten er samme variabel. Videre gir anova-analyser av de to variantene av variabelen også de samme faktorene, men p-verdien blir noe høyere med klausus som målestokk, $p = 0{,}061$, og resultatet dermed ikke signifikant. Figur 9-8 viser i hvilken grad tendensene likevel er de samme som i figur 9-7, om enn altså ikke like sterke.



Figur 9-8: Effekten av skriveverktøy på leksikalsk tetthet målt som antall leksikalske ord per klausus. Til venstre differanse i leksikalsk tetthet fordelt etter skriveferdighet og tekstlengdeforskjell. I midten leksikalsk tetthet fordelt etter skriveverktøy og tekstlengdeforskjell, men bare for sterke elever. Til høyre et spredningsdiagram som viser sammenhengen mellom tekstlengdeforskjell, leksikalsk tetthet og skriveoppgave. Diagrammene svarer til diagrammene i figur 9-7 for antall leksikalske ord per løpeord.

For det andre er det usikkerhet om definisjonen av leksikalske ord. Jeg har i opptellingen av leksikalske ord foretatt en rekke diskuterbare valg i kategorisk inndeling av klasser som i realiteten er prototypiske, og jeg kunne ha valgt både mer og mindre omfattende definisjoner. Jeg har til en viss grad eksperimentert med å variere definisjonen av leksikalske ord, og andre definisjoner gir i noen tilfeller litt andre minimale anova-modeller, selv om tendensene generelt er de samme. Imidlertid er det klart fra framstillingen over at det kan knyttes en god del usikkerhet til det resultatet som presenteres.

For det tredje, og delvis relatert til diskusjonen over om definisjonen, er det problematisk å dikotomisere en variabel som jeg har vist har så fremtredende graduelle eller prototypiske egenskaper. Som Halliday (1989, s. 65) er inne på, kunne de leksikalske ordene vektes etter frekvens, slik at sjeldnere ord fikk tillagt mer vekt i analysen og ble regnet som enda "mer leksikalske" fordi de er bærere av mer informasjon enn mer frekvente leksikalske ord. I 9.3 nedenfor utforsker jeg kort to ulike typer muligheter for en slik tilnærming.

9.3 Leksikalsk spesifisitet

9.3.1 Leksikalsk sofistikerthet og leksikalsk originalitet

Vekting av leksikalitet etter bruksfrekvens har en klar forbindelse med et informasjonsteoretisk perspektiv på tekst, ved at man ser på sjeldne ord som bærere av mer informasjon enn frekvente ord. Men sjeldne ord er gjerne bærere av mer informasjon ikke bare rent informasjonsteoretisk ut fra sin sjeldenhet, men også fordi de gjerne er semantisk mer spesifikke, intensjonelt og/eller ekstensjonelt. På den måten øker sjeldne ord informasjonstettheten i en tekst uten at dette gir seg utslag i økt leksikalsk tetthet.

Når man skal klassifisere eller gradere ord som sjeldne eller frekvente, er det et relevant spørsmål hva slags tekstsamling man skal legge til grunn som referansekorpus. Ideelt kunne man tenke seg å basere seg på all tekst som er tilgjengelig i språksamfunnet, men det er

selvfølgelig umulig å gjennomføre både praktisk og prinsipielt. Det er dessuten kanskje heller ikke prinsipielt særlig valid, om man beholder det informasjonsteoretiske perspektivet. Frekvensen til et ord i en tekst sier noe om informasjonsinnholdet bare innenfor de sjangermessige eller funksjonelle rammene som teksten opererer og fungerer innenfor. Mer relevant er det derfor å ta utgangspunkt i en samling av tekster som sjangermessig er så nært beslektet med studieobjektet som mulig. I dette tilfellet kunne det for eksempel være en samling av samfunnsdebatterende tekster.

Innenfor rammene av dette prosjektet måtte jeg ta utgangspunkt i en eksisterende frekvensordliste, og da var frekvensordlista fra Oslo-korpuset (Tekstlaboratoriet, 2010) det mest nærliggende valget, nærmere bestemt de 10 000 mest frekvente ordformene fra avistekster. Det er en ulempe at lista er basert på ordformer og ikke lemmaformer eller leksemer. Det er dessuten en ulempe at den ikke er lengre enn 10 000 ordformer. Den dekker uansett omtrent 86 % av de leksikalske løpeordene i elevtekstkorpuset, men bare 60 % av de leksikalske ordformtypene.

En alternativ tilnærming er å la elevtekstkorpuset selv danne grunnlaget for frekvenslista og det informasjonsteoretiske universet. På den måten kan man bygge frekvenslista på lemmaformene, og man er garantert at alle lemmaformene i korpuset er til stede i frekvenslistene. En slik fremgangsmåte kan selvfølgelig også ha visse validitetsmessige ulemper, som jeg skal komme tilbake til.

Den første av disse metodene er relatert til det Linnarud (1986, s. 45) kaller *lexical sophistication* og Monsen (2008, s. 44) kaller leksikalsk frekvensprofil. Den andre metoden er relatert til det Linnarud (s. 44) kaller *lexical individuality*. De nøyaktige utregningsmetodene jeg bruker, avviker imidlertid fra alle disse nevnte metodene. Jeg kaller den aviskorpusbaserte indeksen *leksikalsk sofistikerthet* og den elevkorpusbaserte indeksen *leksikalsk originalitet*.

9.3.1.1 Utregning

Jeg har utviklet en formel som jeg bruker til å regne ut en frekvensindeks som jeg kaller *lfi* (for logaritmisk frekvensindeks). *Lfi* for hver tekst er bestemt av følgende formel:

$$lfi = \frac{\sum \frac{-\log\left(\frac{k_i}{N}\right)}{\log(N)}}{n}$$

der n er antall leksikalske løpeord i teksten, N er antall ordformer totalt i referansekorpuset som ligger til grunn for frekvenslista, 9 520 502 i aviskorpuset og 59 754 i elevkorpuset, og k_i er antall forekomster i referansekorpuset av den leksikalske ordformen i. Ordformer som ikke forekommer i frekvenslista, får tildelt en stipulert k-verdi som er halvparten av den laveste k-verdien i frekvenslista. I avislista blir den stipulerte k-verdien 34; når elevtekstkorpuset brukes til å generere frekvenslista, er det selvfølgelig ikke aktuelt med en stipulert k-verdi.

Brøken k_i/N angir hvor sjelden ordformen er i referansekorpuset. Den mest frekvente utvilsomt leksikalske ordformen i aviskorpuset er \sier\, med $k/N \approx 0{,}0030$. Den minst frekvente ordformen i den 10000 ord lange frekvensordlista fra aviskorpuset er \speilet\, med $k = 67$, $k/N \approx 0{,}0000070$. Den negative logaritmeverdien av k_i/N blir lavere jo sjeldnere ordformen er; for \sier\ er den $\log(k_{sier}/N) \approx -5{,}8$, mens $\log(k_{speilet}/N) \approx -11{,}9$. Negering og divisjon med $\log(N)$ sikrer en verdi mellom 0 og 1 for hver ordform. Ordformer som forekommer bare 1 gang i referansekorpuset, får $lfi_{max} = 1$, mens den teoretiske minimumsverdien for en ordform som fyller hver plass i korpuset, er $lfi_{min} = 0$. Høye *lfi*-verdier representerer altså sjeldenhet eller lav frekvens.

For hver leksikalske ordform *i* i en tekst beregnes en lfi_i, og for hver tekst beregnes *lfi* som middelverdien av de leksikalske løpeordene. En høyere gjennomsnittlig *lfi* for en tekst representerer dermed mer bruk av sjeldne (leksikalske) ord.

9.3.1.2 Deskriptiv analyse

De konkrete verdiene av begge de to variantene av *lfi* avhenger selvfølgelig av egenskapene til referansekorpuset som ligger til grunn for beregningene. Formålet med variabelen er å studere forskjellene mellom verdier, mens de konkrete verdiene er lite relevante i seg selv – originalitetsverdiene i enda mindre grad enn sofistikerthetsverdiene. Jeg oppgir dem likevel her som grunnlag for analyse og diskusjon.

Tabell 9-9: Nøkkeltall for leksikalsk sofistikerthet, *lfi* basert på aviskorpus

	middel	median	sd	min	maks
Total	**0,588**	**0,587**	**0,013**	**0,556**	**0,622**
Hånd	0,589	0,587	0,013	0,567	0,620
Tast	0,587	0,588	0,014	0,556	0,622
Middels	0,587	0,586	0,014	0,556	0,622
Sterk	0,590	0,589	0,013	0,568	0,614
Gutt	0,592	0,590	0,013	0,567	0,622
Jente	0,585	0,584	0,012	0,556	0,614

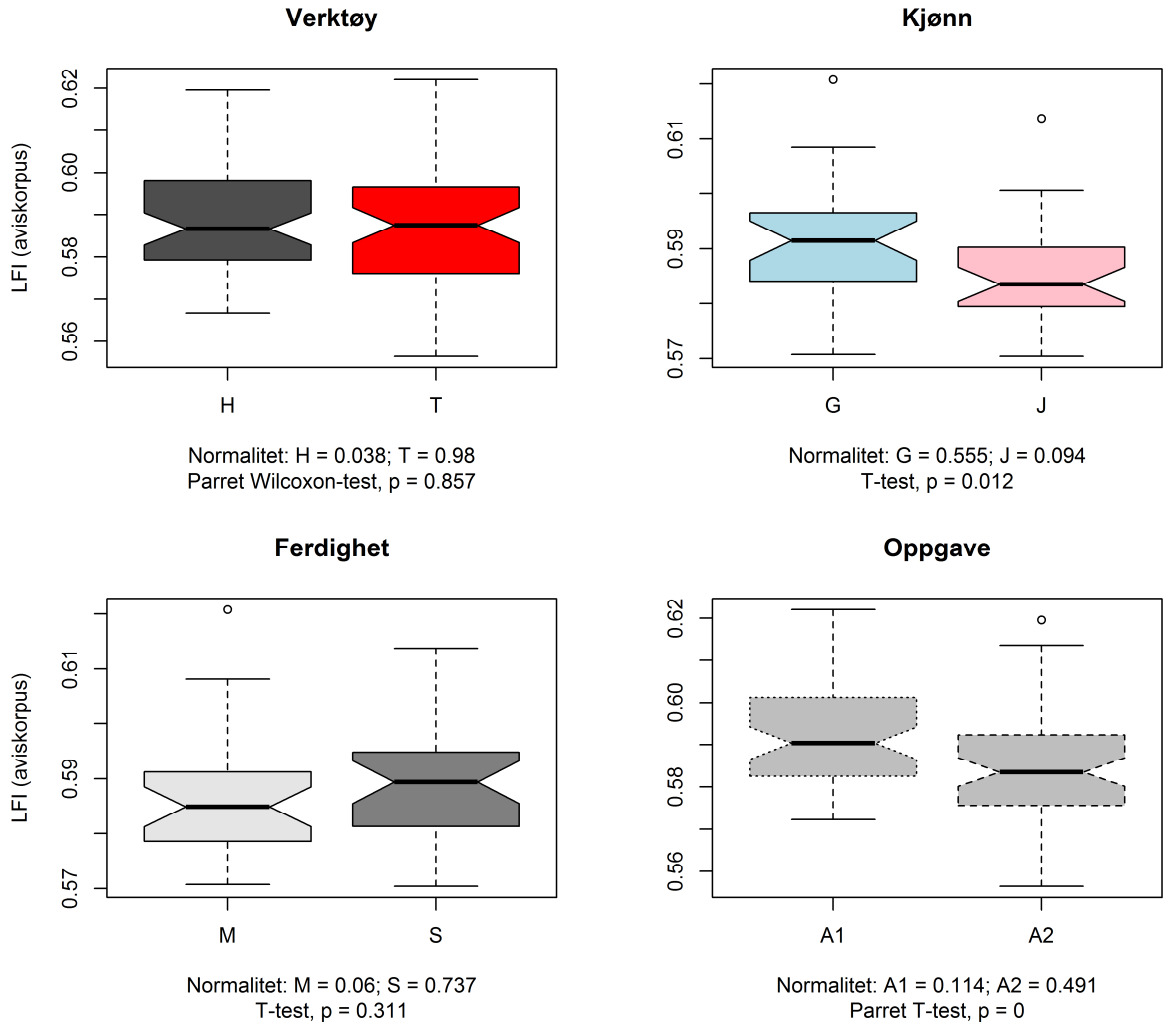
Median- og middelverdiene for leksikalsk sofistikerthet ligger rundt 0,59, med standardavvik på ca. 0,013. Utvalget er normalfordelt, men ikke alle de relevante delutvalgene er det, noe som kan ha konsekvenser for analysen under.

Tabell 9-10: Nøkkeltall for leksikalsk originalitet, *lfi* basert på elevkorpus

	middel	median	sd	min	maks
Total	**0,633**	**0,634**	**0,038**	**0,532**	**0,729**
Hånd	0,634	0,631	0,036	0,561	0,717
Tast	0,632	0,637	0,039	0,532	0,729
Middels	0,627	0,628	0,034	0,561	0,712
Sterk	0,639	0,646	0,040	0,532	0,729
Gutt	0,645	0,647	0,032	0,561	0,729
Jente	0,621	0,624	0,039	0,532	0,723

Median- og middelverdiene for leksikalsk originalitet er noe høyere enn for leksikalsk sofistikerthet. Dette er ikke overraskende, ettersom elevkorpuset er såpass mye mindre enn aviskorpuset, og et mindre korpus vil resultere i en mindre verdi for log(*N*) i kvotienten i formelen for *lfi*. Middelverdiene ligger rundt 0,63, mens standardavviket er på rundt 0,037; det er altså vesentlig mer variasjon i originalitetsvariabelen enn i sofistikerthetsvariabelen.

Figur 9-9 og figur 9-10 nedenfor viser at guttene har høyere *lfi*, uansett om frekvenslistene er basert på avis- eller elevkorpuset, mens det ikke er noen effekter hverken av skriveverktøy eller skriveferdighet.

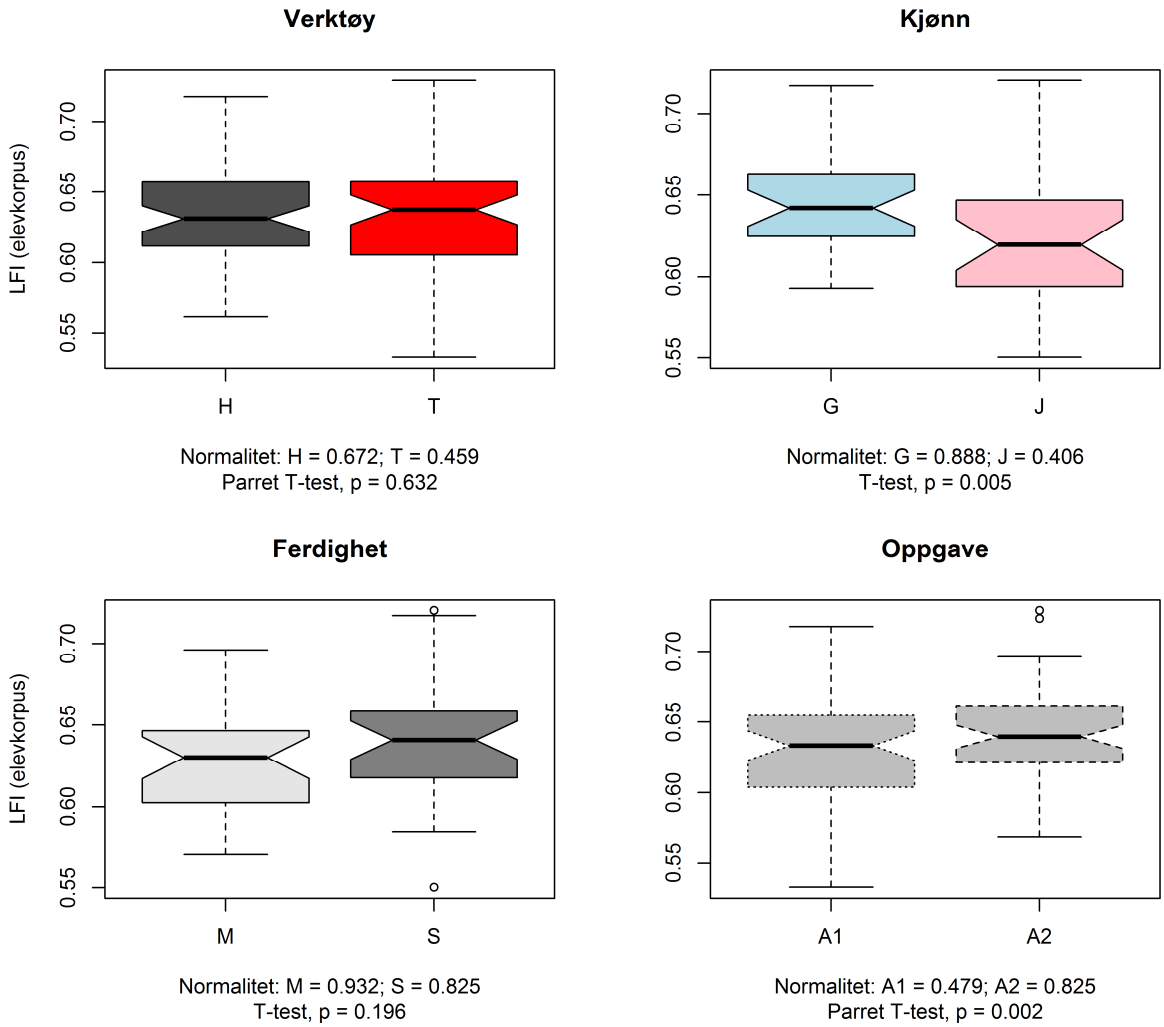


Figur 9-9: Leksikalsk sofistikerthet, altså *lfi* basert på avistekstkorpuset

Imidlertid viser sofistikerthetsindeksen og originalitetsindeksen motsatt effekt når det gjelder oppgavetekst. Med utgangspunkt i aviskorpuset har "Bøker eller data"-tekstene høyest *lfi*, mens "Ungdomsfylla" har høyest *lfi* når indeksen tar utgangspunkt i elevtekstkorpuset. Jeg vil ikke gå for langt i fortolkningen av disse forskjellene, men det virker naturlig at de lengste tekstene, altså "Ungdomsfylla"-tekstene, får høyere originalitetsverdier, rett og slett fordi potensialet for unike korpusord øker etter hvert som tekstenes lengde øker. Figur 9-11

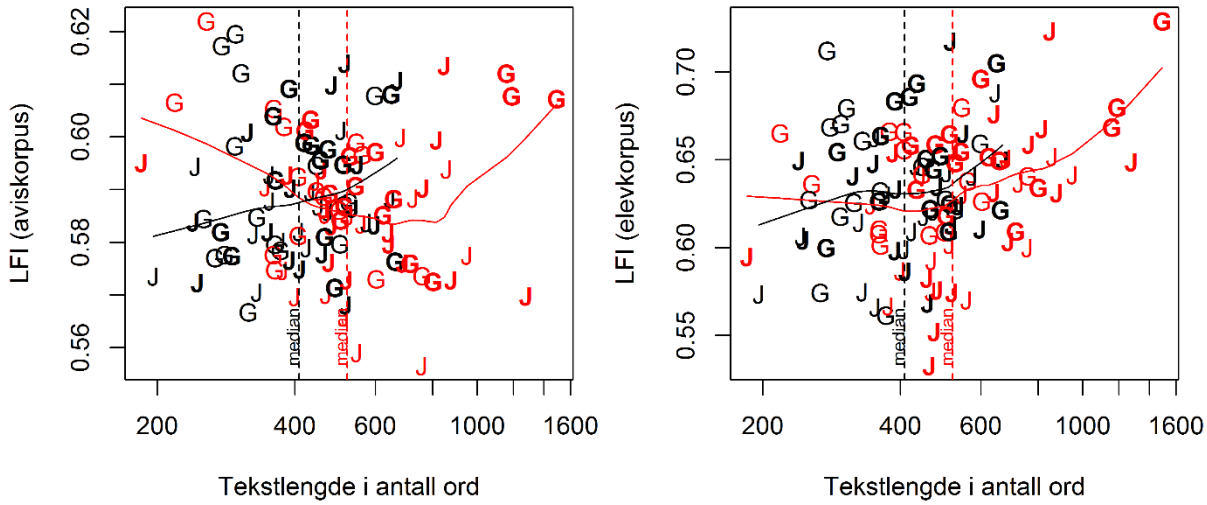
viser da også en generell sammenheng med tekstlengde for leksikalsk originalitet, men ingen slik tendens for sofistikerthet.

Når det gjelder den motsatte tendensen for sofistikerthetsindeksen, altså at "Ungdomsfylla"-tekstene har lavere verdier enn "Bøker eller data"-tekstene, kan man lett tenke seg at dette skriver seg fra at avistekstene i større grad handler om alkoholbruk enn om bøker og pc-bruk, kanskje særlig i og med at alle avistekstene skriver seg fra før 1999. Denne indeksen krever trolig et mer moderne, mer omfattende og mer sofistikert korpus som grunnlag.



Figur 9-10: Leksikalsk originalitet, altså *lfi* med utgangspunkt i elevtekstkorpuset

De to variabelvariantene har som nevnt også ulike tendenser med hensyn til sammenheng med tekstlengde. Leksikalsk originalitet (til høyre i figur 9-11) viser en generell positiv korrelasjon med logaritmisk tekstlengde, R ≈ 0,27, mens leksikalsk sofistikerthet (til venstre) viser ingen generell korrelasjon med logaritmisk tekstlengde, R ≈ –0,02, men derimot tilsynelatende store forskjeller mellom håndtekster og tastetekster. Se diskusjonen i 9.3.1.4 nedenfor.



Figur 9-11: De to varianter av *lfi* og sammenheng med tekstlengde. Til venstre leksikalsk sofistikerthet og til høyre leksikalsk originalitet.

9.3.1.3 *Variansanalyse*

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, antall interaksjonsnivåer begrenset til 2:

(105) `lm(lexD$avislfi~(kjønn+ferdighet+lengde+forskjell)^2)`

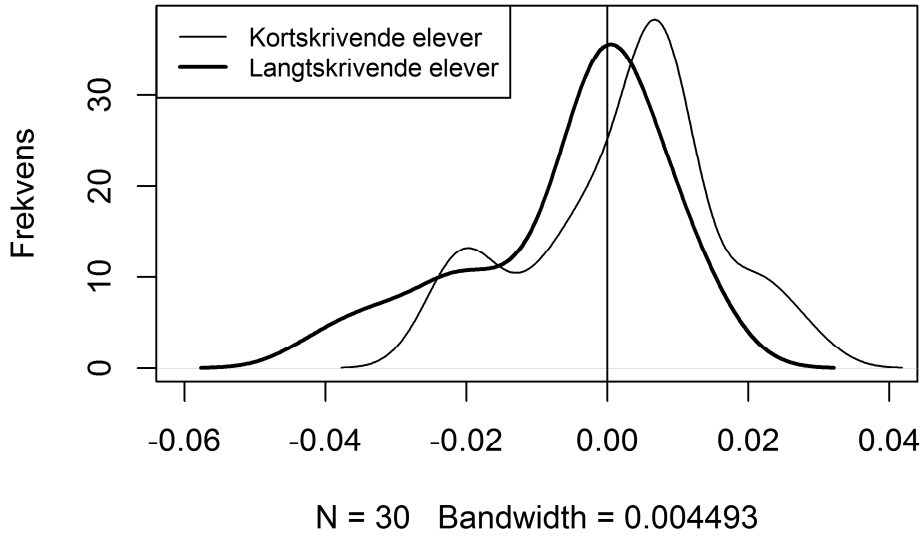
Analysen resulterte i følgende minimale adekvate modell, som er svakt signifikant ($F \approx 4{,}5$, $p < 0{,}05$):

(106)
```
lm(formula = lexD$avislfi ~ lengde)

             Df   Sum Sq   Mean Sq F value Pr(>F)
lengde        1 0.000872 0.0008720   4.544 0.0373 *
Residuals    58 0.011129 0.0001919
---
Multiple R-squared:  0.07266,   Adjusted R-squared:  0.05667
F-statistic: 4.544 on 1 and 58 DF,  p-value: 0,03727
```

I analysen av leksikalsk sofistikerthet er total tekstlengde den eneste signifikante faktoren, $p < 0{,}05$. Diagrammet til venstre i figur 9-13 viser at elever som skriver langt, har en tendens til høyere spesifisitet i håndtekstene, mens elever som skriver kort, har en tendens til lavere spesifisitet i håndtekstene, mens tastetekstene har omtrent de samme verdiene for elever som skriver kort og elever som skriver langt.

`Gvlma` (se 7.2.2.4) viser at utvalget er skjevt, $p = 0{,}046$ (se appendiks A4), så vi må ta et visst forbehold ved resultatet. Skjevheten er imidlertid ikke dramatisk, og den går i samme retning for begge de aktuelle delutvalgene (figur 9-12). Dessuten er utvalgene ganske store ($N = 30$) og like store. I lys av diskusjonen rundt t-testens robusthet (7.2.3.1) kan man etter min mening feste ganske stor lit til dette resultatet.



Figur 9-12: Tetthetskurver for differanser i aviskorpusbasert *lfi*. Diagrammet viser at variabelen er lett venstreskjev både for elever som skriver korte tekster, og for elever som skriver lange tekster.

Når analysen tar utgangspunkt i frekvenslistene fra elevtekstkorpuset, altså leksikalsk originalitet, blir resultatet derimot et ganske annet. Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, med antall interaksjonsnivåer begrenset til 2:

(107)
```
lm(lexD$elevlfi ~ (kjønn+ferdighet+lengde+forskjell)^2)
```

Den maksimale modellen over kan reduseres til den minimale adekvate modellen under:

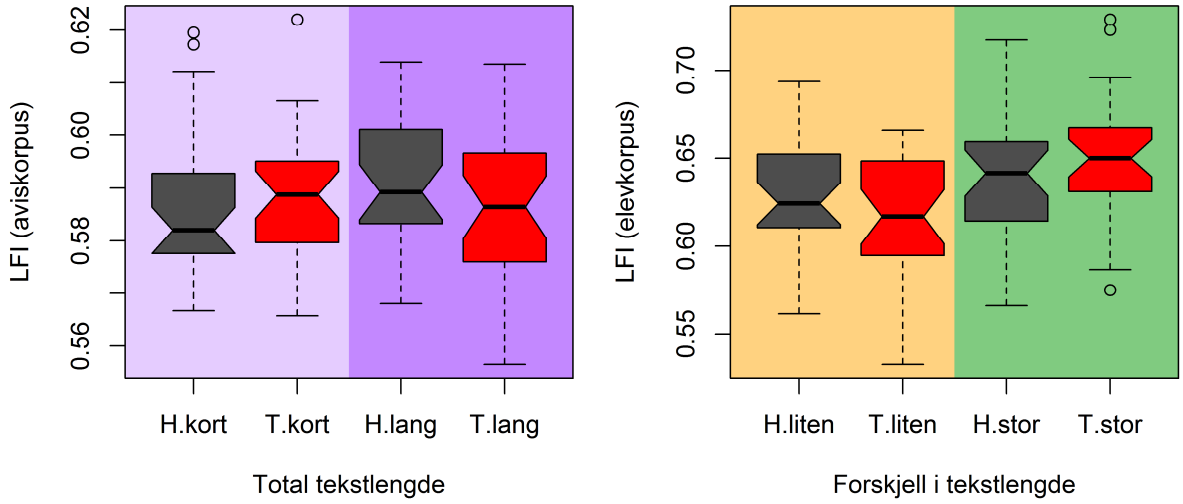
(108)
```
lm(formula = lexD$elevlfi ~ forskjell)

             Df  Sum Sq  Mean Sq F value Pr(>F)
forskjell     1 0.00569 0.005693   6.043  0.017 *
Residuals    58 0.05464 0.000942
----
Multiple R-squared:  0.09436,   Adjusted R-squared:  0.07875
F-statistic: 6.043 on 1 and 58 DF,  p-value: 0.01697
```

Da er *forskjell* i tekstlengde den eneste signifikante faktoren ($F \approx 6{,}0$, $p < 0{,}05$), og de elevene som skriver mye lengre på tastatur, har også tendens til høyere spesifisitet på tastatur, mens det motsatte er tilfellet for elevene som ikke skriver mye lengre på tastatur. Dette virker i utgangspunktet litt kontraintuitivt og ser dessuten ut til å være den motsatte effekten av den vi så for leksikalsk tetthet når det gjaldt sterke elever (se 9.2.4 ovenfor). Imidlertid er det ikke så unaturlig ut fra diskusjonen omkring sammenhengen mellom tekstlengde og leksikalsk originalitet. Dessuten er det en effekt som ligner den vi skal se for TTR-baserte mål i figur 10-18 på side 191, 10.3.5 og 10.4.5.3. Se også diskusjonen i 9.3.1.4 nedenfor.

`Gvlma` (se 7.2.2.4) rapporterer at premissene er oppfylt (se appendiks A4).

Figur 9-13 oppsummerer effekten av de to variantene av *lfi*.



Figur 9-13: Diagrammer som viser interaksjoner mellom skriveverktøy og prediktorer for de to variantene av *lfi*. Til venstre leksikalsk sofistikerthet med total tekstlengde som signifikant prediktor. Til høyre leksikalsk originalitet med forskjell i tekstlengde som signifikant prediktor.

9.3.1.4 Oppsummering og diskusjon

Det er vanskelig å tolke disse resultatene. For frekvensanalysen basert på lista fra elevtekstkorpuset er det naturlig at tallene øker med tekstlengde. Siden alle tekstene handler om de samme to temaene, er det naturlig at flertallet av tekster av normal lengde deler mye av ordforrådet, mens de riktig lange tekstene i større grad kommer inn på momenter eller ordbruk som skiller seg fra resten. Det innebærer ikke nødvendigvis at de bruker ord som er lite frekvente *i språket* eller i sjangeren generelt.

At man bruker ord som ikke er mye brukt av de andre tekstene, kan innebære at man skriver mer spesifikt eller nyansert, bruker andre typer metaforer, eller at man er inne på flere eller andre momenter enn de andre elevene. Men det kan også skrive seg fra at man ikke holder seg til saken, men kommer inn på irrelevante sidespor. Kombinasjonen av alle disse mulige årsaksforholdene gjør at en variabel som er knyttet til hvor sjeldne ord er *i korpuset*, er vanskelig å tolke. Det samme forbeholdet gjelder analysene basert på frekvenslista fra aviskorpuset, om enn kanskje ikke i like stor grad. I alle fall er det grunn til å stille spørsmål ved om begge teksttemaene er like bredt representert i aviskorpuset, og om en eventuell skjevfordeling kan interagere med ulikheten i tekstlengde mellom de to oppgavene i elevkorpuset.

Begge resultatene er knyttet til tekstlengde, og interaksjonene mellom tekstlengde og oppgavetypene gjør dem vanskelige å tolke. At tastetekstene om "Ungdomsfylla" (oppgave A2) er så mye lengre enn de andre delutvalgene, gjør som ellers i denne studien at en effekt av tekstlengde i realiteten kan være en effekt av oppgavetypen. Men i dette tilfellet forstyrres også frekvenslistene av at A2-tekstene er lengre og inneholder flere ulike lemmaformer. Vokabularet i de to oppgavene er ikke likt, så frekvenslistene fra elevkorpuset inneholder

flere lemmaformer som er relevante for A2, slik at originalitetsverdiene for A1 kan bli kunstig høye. Akkurat denne effekten av ulik tekstlengde kunne kanskje ha blitt nøytralisert ved å legge de to delkorpusene separat til grunn for frekvenslistene i analysen.

Tabell 9-11: De 30 mest frekvente leksikalske lemmaformer fra hvert delkorpus

A1 (Bøker eller data)		**A2 (Ungdomsfylla)**	
gutt	556	drikke	468
lese	554	forelder	423
jente	518	ungdom	386
bøk	474	alkohol	375
datum	371	få	283
drive	244	ha	282
ha	175	barn	259
gjøre	170	fest	247
sitte	158	gjøre	225
bruke	156	komme	189
si	140	bli	172
påstand	122	si	148
dag	108	tro	133
data	107	kjøpe	121
pc	104	øl	110
komme	102	år	98
tro	100	litt	97
bli	99	svar	91
ungdom	97	vite	91
tid	96	dag	90
hel	95	enkelt	90
bok	94	god	90
like	93	gi	86
få	75	negativ	82
spille	71	stor	81
se	70	mene	80
god	63	redaktør	76
feil	59	full	74
ta	58	hel	74
mene	57	se	74

Tabell 9-11 ovenfor viser i hvilken grad det leksikalske ordtilfanget i de to oppgavene er ulikt. Samtidig kan man se at antall forekomster generelt synker raskere i A1-lista enn i A2-lista, noe som skriver seg fra kombinasjonen av den større totale ordmengden og bredden i ordtilfanget i A2. Det er 1634 forskjellige lemmaformer i A1-tekstene, 1758 i A2-tekstene.

Også andre delvis praktiske forhold vanskeliggjør tolkningen av disse variablene, først og fremst knyttet til begrensninger ved aviskorpuset, som er diskutert over. Disse forholdene gjør det naturlig å reise spørsmål rundt validiteten ved resultatene.

Linnarud (1986) operasjonaliserer originalitet, eller individualitet, på en annen måte enn meg; hun teller hvor mange leksikalske ord i hver tekst som er unike for korpuset, og dividerer på antall leksikalske løpeord i teksten. Dette stiller strengere krav til individualitet eller originalitet. Jeg har også forsøkt Linnaruds fremgangsmåte, og den resulterer i en anova-modell uten signifikante resultater.

9.3.2 Ordlengde i leksikalske ord

Som jeg er inne på i kapittel 9.1, kan ordlengde reflektere flere egenskaper ved ord og tekst. I tillegg til at det er vesentlig forskjell i gjennomsnittlig ordlengde mellom grammatiske og leksikalske ord, kan lange ord signalisere høyere presisjon, mer nyansering, variasjon og abstraksjon og mer integrert språk.

Imidlertid gjør den temmelig sterke korrelasjonen mellom leksikalsk tetthet og ordlengde, $R \approx 0{,}68$, at de andre tekstegenskapene som kan knyttes til ordlengde, kan komme i bakgrunnen. Ved å undersøke gjennomsnittlig ordlengde bare blant de leksikalske ordene kan man kanskje greie å isolere andre egenskaper enn leksikalsk tetthet.

I motsetning til flere av de andre leksikalske variablene i denne avhandlingen er antall bokstaver per leksikalske enhet et svært intuitivt mål som ikke krever noen transformering. Som mål på presisjon, nyansering og andre typer egenskaper relatert til informasjonsteoretisk innhold ville det være mest naturlig å bruke lemmaformens lengde som måleenhet. For å beholde et best mulig sammenligningsgrunnlag med variabelen for gjennomsnittlig ordformlengde i 9.1 har jeg imidlertid valgt å beholde ordformen som måleenhet også for lengden av de leksikalske ordene.

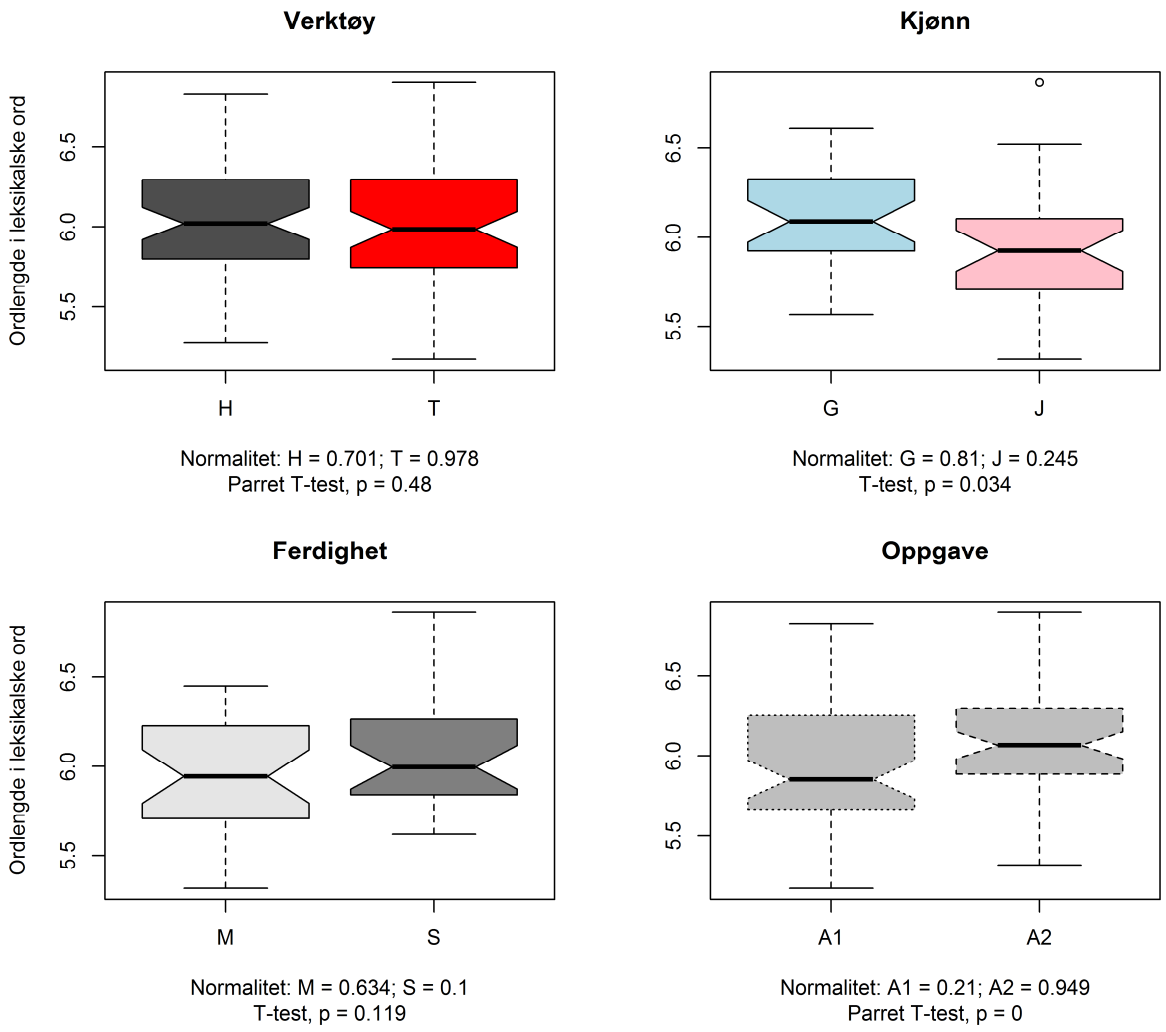
9.3.2.1 Deskriptiv analyse

Blant de leksikalske ordformene er middelverdien for ordlengde i elevtekstkorpuset 6,0 bokstaver, med standardavvik rundt 0,35, som vist i tabell 9-12. Utvalget og de relevante delutvalgene er normalfordelt.

Tabell 9-12: Nøkkeltall for gjennomsnittlig leksikalsk ordlengde

	middel	median	sd	min	maks
Total	**6,02**	**6,01**	**0,35**	**5,17**	**6,90**
Hånd	6,04	6,02	0,33	5,27	6,83
Tast	6,01	5,98	0,38	5,17	6,90
Middels	5,96	6,00	0,34	5,17	6,53
Sterk	6,09	6,01	0,36	5,32	6,90
Gutt	6,11	6,07	0,30	5,55	6,71
Jente	5,94	5,90	0,38	5,17	6,90

Figur 9-14 nedenfor viser at guttene har vesentlig høyere leksikalsk ordlengde enn jentene, $d \approx 0{,}57$, og at "Ungdomsfylla"-tekstene også har høyere verdier enn "Bøker eller data"-tekstene, $d \approx 0{,}46$.



Figur 9-14: Gjennomsnittlig ordlengde for leksikalske ord

Det er ingen tendens til korrelasjon med tekstlengde, R ≈ 0,04. Korrelasjonen mellom hånd- og tastetekster er lik for gutter og jenter og ganske sterk, R ≈ 0,51, altså noe lavere enn den er for generell ordlengde (se figur 9-4).

9.3.2.2 Variansanalyse

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, med antall interaksjonsnivåer begrenset til 2:

(109) `lm(lexD$leksordlengde ~ (kjønn+ferdighet+lengde+forskjell)^2)`

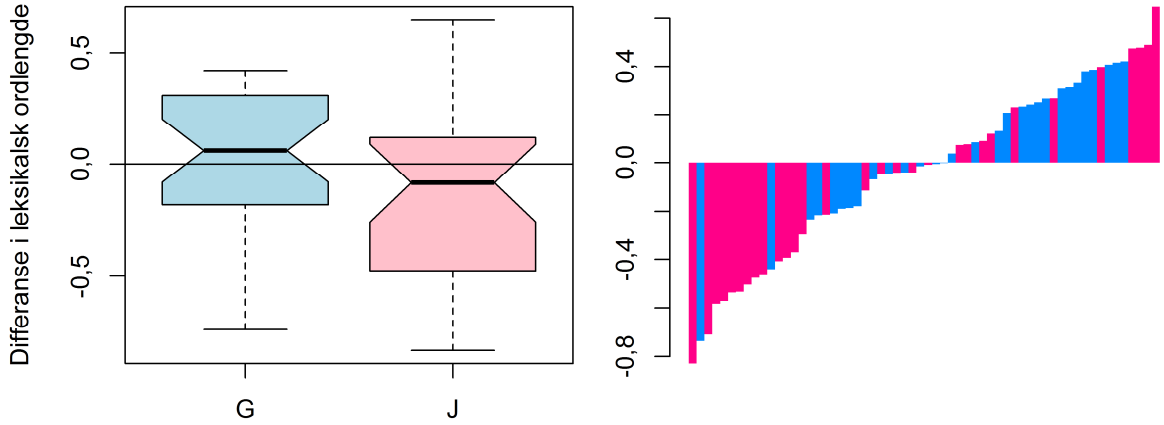
Den maksimale modellen reduseres til følgende minimale adekvate modell (F ≈ 4,4, p < 0,05):

(110) `lm(formula = lexD$leksordlengde ~ kjønn)`

```
            Df Sum Sq Mean Sq F value  Pr(>F)
kjønn        1 0,5338 0,53375  4,4251 0,03976 *
Residuals   58 6,9959 0,12062
---
Multiple R-squared:  0.07089,   Adjusted R-squared:  0.05487
F-statistic: 4.425 on 1 and 58 DF,  p-value: 0.03976
```

`Gvlma` (se 7.2.2.4) viser at premissene for anova-analysen er oppfylt (se appendiks A4).

Det er altså ingen absolutte forskjeller mellom verktøyene, men det er en svakt signifikant forskjell mellom kjønnene ($p < 0,05$) på den måten at guttene har en tendens til høyere leksikalsk ordlengde i tastetekstene, mens jentene har en tendens til lavere, som det går fram av figur 9-15. Forskjellen i differansenes middelverdi er 0,19 bokstaver, som tilsvarer $d \approx 0,55$.



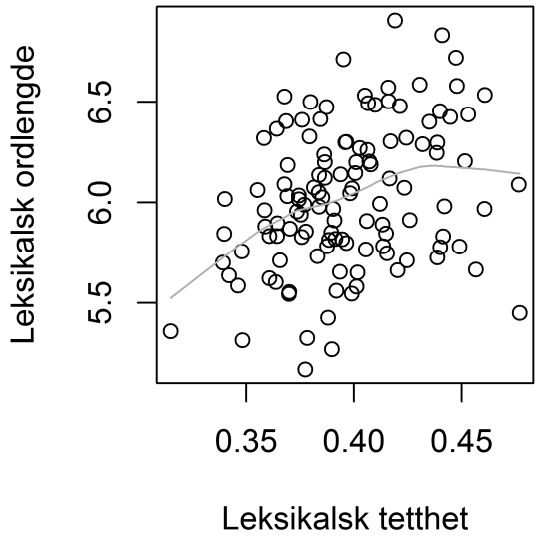
Figur 9-15: Diagrammer som illustrerer resultatet av variansanalyse på leksikalsk ordlengde. Boksdiagrammet til venstre viser de generelle kjønnsforskjellene. Diagrammet til høyre viser forskjellene på individnivå.

Skreddiagrammet over forskjellene i figur 9-15 ovenfor viser at jentene dominerer det nedre sjiktet blant de som har kortere leksikalske ord i tastetekstene, og forskjellene er ganske store; over 0,8 bokstaver i gjennomsnitt over en hel tekst er en dramatisk forskjell, riktignok noe avhengig av tekstenes lengde. Men diagrammet viser også at 4 jenter har de mest ekstreme verdiene på den andre siden, noe som bidrar til at de generelle, gjennomsnittlige kjønnsforskjellene ikke er større. Blant tastetekstene er 14 av de 15 laveste verdiene fra tekster skrevet av jenter (ikke vist i diagram).

9.3.2.3 Oppsummering og diskusjon

Effekten av skriveverktøyet på variabelen leksikalsk ordlengde ligner ganske mye på effekten på generell ordlengde (9.1.3); gutter skriver mer planlagt på tastatur, mens jenter skriver mer spontant på tastatur. At tilsvarende kjønnseffekt ikke finnes for leksikalsk tetthet (9.2.4), tilsier at endringen ordlengde ikke først og fremst kommer av påvirkning på den leksikalske tettheten i tekstene.

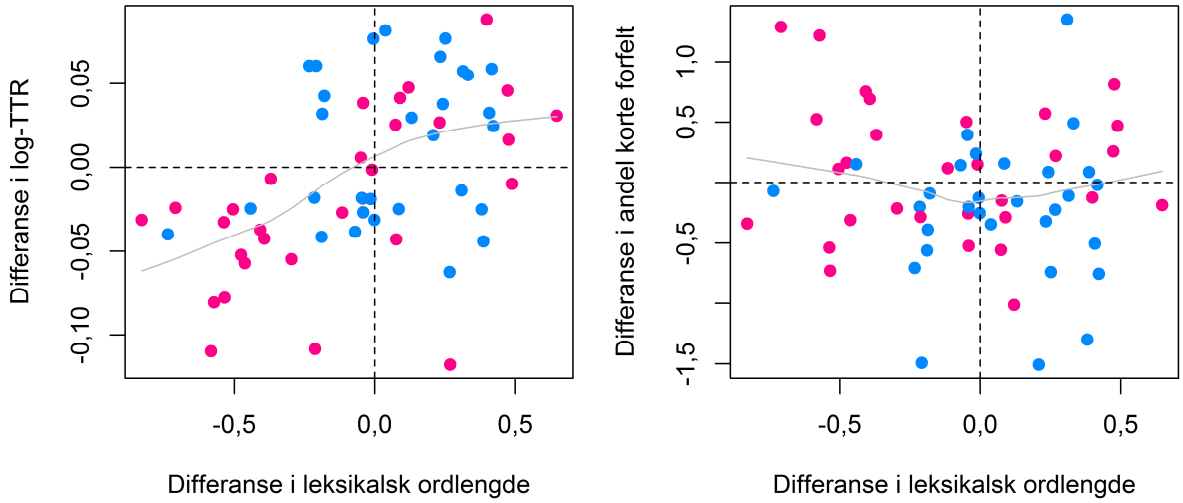
Selv om korrelasjonen mellom gjennomsnittlig ordlengde og leksikalsk tetthet reduseres betraktelig når man regner ordlengde bare av de leksikalske ordene, er det fortsatt en klar sammenheng, R ≈ 0,33,[18] særlig for tekstene med lavere verdier av leksikalsk tetthet, se figur 9-16. Denne sammenhengen kan ikke skrive seg fra den leksikalske tettheten i seg selv, så det må bety at elever som bruker høyere andel av leksikalske ord, dessuten bruker lengre leksikalske ord, altså skriver med et mer presist eller nyansert ordforråd. Imidlertid kan, som jeg har argumentert for over, lengre ord også reflektere tekstegenskaper som ikke direkte har med presisjon eller nyansering å gjøre. For eksempel representerer nominaliseringer en type integrerthet som vil kunne øke både den leksikalske ordlengden og den leksikalske tettheten uten at presisjonen nødvendigvis blir høyere. Korrelasjonen mellom disse to variablene kan med andre ord skrive seg fra egenskapen integrerthet alene, men mer trolig er det en kombinasjon av flere egenskaper.



Figur 9-16: Sammenheng mellom leksikalsk tetthet og leksikalsk ordlengde

At gutter har høyere verdier for denne variabelen i tastetekster, og særlig at jenter har lavere verdier, stemmer godt med resultatene for flere av de andre variablene i studien, for eksempel log-TTR (se 10.6) og andel t-enheter med ett ord i forfelt (se 11.3.2). Den viser at et flertall av guttene bruker verktøyet til å forsterke tekstens planlagte eller integrerte egenskaper, mens et flertall av jentene bruker verktøyet til å forsterke tekstens mer spontane egenskaper. Blant jentene ser det ut til at en stor del av forskjellen skyldes en håndfull jenter med spesielt lave verdier i tastetekstene.

[18] Korrelasjonen er her regnet ut på samtlige 120 tekster i korpuset. Ettersom det bare er 60 elever, innebærer det at verdiene ikke er uavhengige, og derfor oppgir jeg heller ikke p-verdi.



Figur 9-17: Ordlengde i leksikalske ord: differanseverdienes korrelasjon med differanseverdiene for log-TTR og andel korte forfelt (logit-transformert)

Figur 9-17 viser imidlertid at det ikke nødvendigvis er *de samme* jentene som velger mer spontane uttrykk i alle variabler.

9.4 Oppsummering og diskusjon

Jeg har i dette kapitlet utforsket og analysert ulike mål for informasjonell tetthet i tekster, spesielt med tanke på å avdekke forskjeller mellom tastede og håndskrevne tekster. Dette delkapitlet gir en kort oppsummering, mens alle de leksikalske variablene sees i sammenheng i kapittel 10.6.

Den første variabelen som ble behandlet, var tekstens gjennomsnittlige ordlengde, målt i antall bokstaver per ordform. Jeg argumenterte for at høyere gjennomsnittlig ordlengde kan være uttrykk for et knippe ulike språklige egenskaper som det kan være vanskelig å måle på andre måter. Dessuten påviste jeg sterk korrelasjon mellom gjennomsnittlig ordlengde og leksikalsk tetthet. Variansanalysen avdekket en kjønnseffekt på effekten av skriveverktøy; gutter skriver med høyere gjennomsnittlig ordlengde på tastatur enn for hånd, mens jenter skriver med lavere gjennomsnittlig ordlengde på tastatur enn for hånd. Analysen avdekket ingen andre effekter, og heller ingen overordnet effekt av skriveverktøy på elevgruppen som helhet. I Bibers og Hallidays perspektiver er det naturlig å relatere høyere gjennomsnittlig ordlengde til planlagt, redigert, skriftlig språkbruk og lavere gjennomsnittlig ordlengde til spontan, muntlig språkbruk.

Den andre variabelen som ble behandlet, var tekstens leksikalske tetthet, altså antall leksikalske ord dividert på antall ord totalt i teksten. Variansanalysen gav som signifikant resultat en interaksjonseffekt mellom skriveferdighet og tekstlengdeforskjell, slik at sterke elever med liten tekstlengdeforskjell har høyere leksikalsk tetthet i tastetekster enn i håndtekster, mens sterke elever med stor tekstlengdeforskjell har lavere leksikalsk tetthet i tastetekster enn i håndtekster. Det var ingen effekt for middels elever. Litt forenklet kunne dette tolkes som at sterke elever bruker det mer avanserte verktøyet *enten* til å redigere og

planlegge teksten *eller* til å skrive en lengre tekst, men det faktum at forskjellene også gjelder håndtekstenes leksikalske tetthet, gjør denne fortolkningen litt problematisk å holde fast ved. Kanskje burde sammenhengen mellom håndtekstenes lengde og leksikalsk tetthet ha vært trukket inn for å forstå dette resultatet bedre. Det er dessuten flere både teoretiske og praktiske problemer knyttet til begrepet leksikalsk tetthet som gjør denne variabelen noe problematisk å bruke i statistiske språktrekkanalyser.

En mangel ved leksikalsk tetthet som operasjonalisering av informasjonell tetthet er at dikotomisering av ordtilfanget i leksikalske og grammatiske ord ikke fanger opp ulik grad av spesifisitet eller informasjonelt innhold i forskjellige leksikalske ord. Den siste delen av kapitlet utforsker ulike måter å bøte på denne mangelen på. Jeg presenterer et frekvensbasert mål som jeg kaller *lfi*, og analyserer to versjoner av en slik variabel basert på to ulike typer referansekorpus. Begge versjonene synes å ha såpass store ulemper knyttet til seg at de ikke er særlig aktuelle å bruke i denne undersøkelsen, og jeg bruker ikke disse resultatene i den videre diskusjonen.

Deretter måler jeg gjennomsnittlig ordlengde bare blant de ordene jeg regner som leksikalske, og en variansanalyse av denne variabelen gir samme effekt som for gjennomsnittlig ordlengde generelt, altså en tendens til at guttene skriver mer planlagt, redigert og skriftlig med tastatur enn for hånd, mens jentene har den motsatte effekten. Leksikalsk ordlengde synes å fange opp andre og potensielt interessante egenskaper ved tekstene enn leksikalsk tetthet, og korrelasjonen med leksikalsk tetthet sannsynliggjør dessuten at gjennomsnittlig ordlengde generelt er en variabel som i tillegg til å ha den fordelen at den er svært rask å beregne, gjenspeiler relevante tekstlige egenskaper utover leksikalsk tetthet.

Av de variablene som er utforsket i dette kapitlet, ser jeg gjennomsnittlig ordlengde, leksikalsk tetthet og gjennomsnittlig ordlengde i leksikalske ord som de mest verdifulle, og disse tre blir diskutert i sammenheng med de andre leksikalske variablene i 10.6 og brukt i prinsipalkomponentanalysen i kapittel 12.

10 Leksikalsk variasjon

Dette delkapitlet fokuserer på leksikalsk variasjon. Det begynner med en drøfting og analyse av frekvenslister (10.1), men fokuset i kapitlet er hovedsakelig på type/eksemplar-forholdstallet, oftest kalt TTR for *type/token ratio*. De grunnleggende prinsippene for og problemene ved TTR blir presentert og diskutert i 10.2. De to påfølgende delkapitlene danner hovedmomentene i kapitlet, nemlig å finne et valid TTR-basert mål ved hjelp av matematisk transformering (10.3) eller segmentbaserte TTR-mål (10.4). Til slutt i kapitlet er et delkapittel om alternative mål på leksikalsk variasjon (10.5), nemlig basert på antall unike ord (10.5.1) og basert på entropi (10.5.2), før en oppsummering og diskusjon av funnene i dette og det foregående kapitlet (10.6).

10.1 Frekvenslister

Frekvensordlister er lister av ord som er sortert etter frekvens i en tekst eller en tekstsamling. Selv om ordformbaserte frekvenslister i visse tilfeller kan være interessant, er ofte frekvenslister basert på leksemer mer relevant. Det er to hovedgrunner til dette. For det første samles antall ordformer av samme leksem, slik at for eksempel frekvensene av bøyningsformene \være\, \er\, \var\ og \vært\ blir samlet til én verdi for leksemet \VÆRE\. For det andre bidrar lemmatiseringen til at en del homografi løses opp. I elevtekststudien er ikke analysene basert på leksemer, men på lemmaformer. Som en konsekvens av dette er ikke all homografi oppløst; homografe lemmaformer vil fortsatt bli regnet som samme "ord" i analysene.

Som alternativ til generelle lister over alle ord kan ordklassespesifikke frekvenslister i noen tilfeller være mer oversiktlige og mer informative. Allmenne lister preges ofte av den skjevfordeling som finnes mellom ulike ordklasser, som man bedre kan presentere på andre måter, mens ordklassespesifikke lister bedre kan belyse ulike egenskaper ved kategorier av ord.

Det er av mindre interesse å lage frekvenslister for hver tekst; slike lister vil i for stor grad være preget av tilfeldig variasjon, ettersom tekstene i elevtekstkorpuset generelt er for korte for et slikt formål. Tekstene er derfor for frekvenslisteformål samlet i to store grupper, én for håndtekster og én for tastetekster, og i lista i tabell 10-1 nedenfor er disse presentert i hver sin kolonne. Lemmaformene er sortert synkende etter frekvens. Tallene er antall forekomster per 1000 løpeord.

Generelle frekvenslister bringer trolig ikke fram særlig mye informasjon som ikke kommer fram gjennom analyse av ordklassefrekvenser og leksikalsk statistikk. Det er likevel en del forfattere, for eksempel Allwood (1998), som studerer generelle frekvensordlister. Tabellen under viser de 30 mest frekvente lemmaformene i elevtekstkorpuset. I frekvenslister som stammer fra tekster om et lite antall emner, vil emnene naturlig nok slå gjennom som frekvente leksikalske ord i frekvenslistene. Listen under viser at substantiver som \gutt\, \jente\, \bok\ og \ungdom\ og verb som \lese\ og \drikke\ har høyere frekvens enn hva som

vil være vanlig for tekstsamlinger generelt. Alle disse leksikalske ordene har helt sentral relevans for tematikken i elevenes skriveoppgaver.[19]

Tabell 10-1: Frekvenslister for lemmaformer. Tallene er antall per 1000 ord.

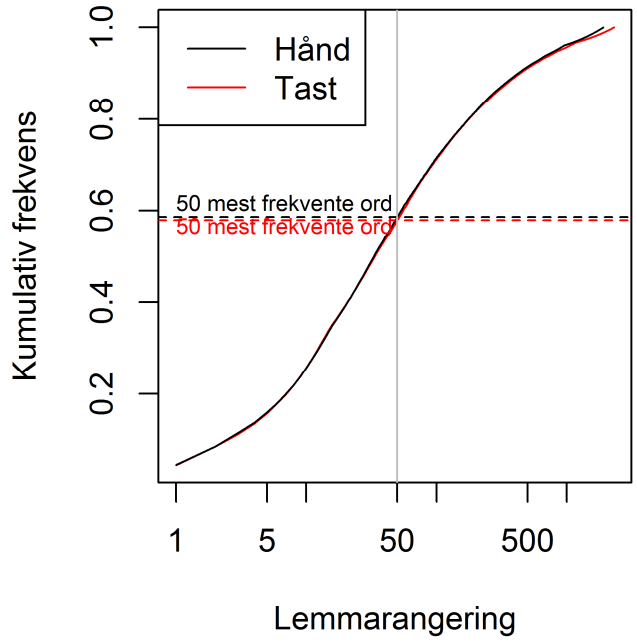
Tast		**Hånd**	
være	4,44	være	4,52
det	4,07	det	3,99
og	2,70	og	3,01
å	2,31	at	2,23
at	2,21	de	2,19
de	2,20	å	2,09
som	1,99	som	1,98
jeg	1,99	på	1,93
ikke	1,97	i	1,89
en	1,87	en	1,84
på	1,86	jeg	1,81
i	1,81	ikke	1,80
ha	1,74	til	1,61
til	1,55	med	1,54
med	1,44	ha	1,45
mye	1,18	men	1,25
men	1,10	for	1,12
kunne	1,03	mye	1,07
for	1,02	gutt	1,06
så	0,94	jente	0,97
mange	0,90	kunne	0,93
gutt	0,88	så	0,91
lese	0,87	drikke	0,90
jente	0,85	mange	0,90
ungdom	0,85	bok	0,88
om	0,82	lese	0,84
bok	0,75	bli	0,81
bli	0,74	forelder	0,78
seg	0,72	ungdom	0,77
drikke	0,71	om	0,76

Allwood bruker frekvensordlistene blant annet til å studere leksikalsk variasjon. Allwood rapporterer at i hans svenske tekster representerer de 50 mest frekvente ordformene i muntlige tekster 52 % av løpeordene, mens i skriftlige tekster representerer de 50 mest frekvente ordformene bare 38 % av løpeordene.

I elevtekstkorpuset er de tilsvarende tallene for lemmaformer 57,9 % for tastetekster og 58,6 % for håndtekster. Det er altså bare en svært liten og trolig tilfeldig forskjell, og den går

[19] I denne typen leksikalske undersøkelser kan det være hensiktsmessig å fjerne alle ord som forekommer i oppgaveteksten (Scott Jarvis, personlig kommunikasjon), men jeg har ikke gjort det i denne undersøkelsen.

dessuten i motsatt retning av hva som kunne forventes ut fra hypotesen (5.1) om at tastetekstene tenderer mot mer spontane trekk. Figur 10-1 viser den kumulative frekvensen for lemmaformer i de to segmentene, og diagrammet demonstrerer at kurvene følger hverandre svært tett. Vi kan riktignok se en liten forskjell nettopp rundt x = 50, men den er svært liten. Forskjellen mellom kurvene når de går mot maksimumsverdien y = 1, skyldes at tastetekstene er lengre og har flere ordtyper enn håndtekstene, og lista for tastetekstene blir derfor lengre enn lista for håndtekstene. Dermed når den svarte kurven opp til y = 1 før den røde.



Figur 10-1: Kumulativ kurve for lemmaformfrekvens. X-aksen er logaritmisk. De stiplede vannrette linjene markerer den kumulative frekvensen for de 50 mest frekvente lemmaformene i hvert delkorpus.

At verdiene er såpass mye høyere enn Allwoods, kan ha flere årsaker. Først og fremst er det en viktig forskjell at Allwoods lister er basert på ordformer, mens elevtekstlistene er basert på lemmaformer. Frekvente leksemer med flere bøyningsformer, som for eksempel \være\ : \er\ : \var\ : \vært\, vil generelt føre til høyere verdier i frekvenslister basert på leksemer eller lemmaformer, selv om Allwood nevner at homografi i enkelte frekvente ordformer vil virke i motsatt retning, for eksempel svensk \att\ som subjunksjon og infinitivsmerke.

For det andre er elevene i elevtekststudien bare ca. 16 år gamle, og de vil med stor sannsynlighet ha lavere frekvens av mer spesifikke ord enn voksne, profesjonelle skribenter som dem i Allwoods studie (Hunt, 1970).

For det tredje påvirkes leksikalske statistikker av tekstlengde, sjanger og emne (Baayen, 2001). Allwood nevner ikke tekstlengde som en faktor, og det går ikke klart fram av artikkelen hvor lange tekstene i hans studie er, men siden deler av det skriftlige tekstutvalget er romaner, kan man gå ut fra at både tekstlengde, sjanger og emne avviker fra elevtekstkorpuset. Dessuten vil en tekstsamling som er satt sammen av ulike teksttyper med

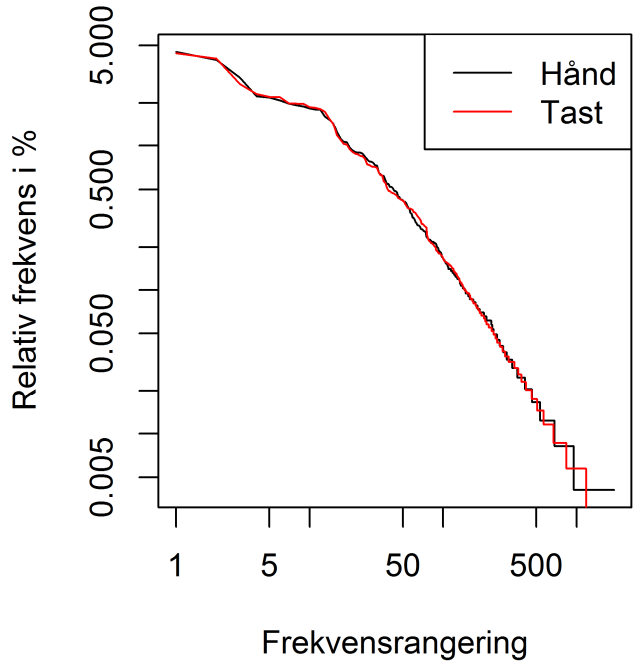
mange ulike emner ha lavere frekvensverdier for de mest frekvente leksikalske ordene enn en tekstsamling som er satt sammen av bare én teksttype med få emner, slik elevtekstkorpuset er.

Til slutt må det nevnes at frekvensmål fra ulike språk aldri kan sammenlignes direkte; ulike språks ulike grammatiske egenskaper vil påvirke resultatene. For eksempel er den mest frekvente ordformen i engelske tekster som regel \the\, som ikke har noen entydig parallell i norsk, men som ofte vil oversettes som en del av ordformen og andre ganger som en av flere alternative determinativformer. Skandinaviske språk som norsk og svensk vil ha mer sammenlignbare resultater.

Den sterke likheten mellom hånd- og tastetekster kommer også fram gjennom en Zipf-kurve for de to delkorpusene. I en Zipf-kurve er hver ordtypes frekvens plottet mot dens nummer i frekvenslista på to logaritmiske akser. Ifølge Zipf (referert av Baayen (2008, s. 226)) vil denne kurven danne en rett linje, og en teksts leksikalske egenskaper kan karakteriseres gjennom linjens stigningstall. Baayen (2008, s. 224-227) tegner Zipf-kurven for romanen *Alice in Wonderland* og viser at kurven ikke er en rett linje, og at det gjennomsnittlige stigningstallet for kurven dessuten varierer med tekstens lengde. Disse to problemene gjør det vanskelig å anvende Zipf-kurvens stigningstall som en tekstkarakteristikk. I det venstre diagrammet i figur 10-3 har jeg som Baayen tegnet Zipf-kurven for ordformene i *Alice in Wonderland*, og diagrammet viser at kurven er konveks for denne romanen.[20]

Figur 10-2 viser at også Zipf-kurvene for de to delkorpusene følger hverandre svært tett, men de danner i likhet med kurven for *Alice in Wonderland* konvekse kurver. Konveksiteten indikerer at vanlige ord – bortsett fra de aller vanligste – har høyere frekvens enn hva Zipfs teorem hevder. De mest frekvente leksikalske ordene befinner seg i dette frekvensbåndet, og det er ikke så overraskende at nettopp barneromaner og korte tekster skrevet av tenåringer i løpet av stramme tidsrammer har en slik egenskap. På bakgrunn av dette er det naturlig å anta at også Zipf-kurvens form må tas med i betraktning dersom den skal kunne fungere som et mål for teksters egenskaper.

[20] For å kunne sammenligne kurvene for de tre tekstene i figuren har jeg tegnet Zipf-kurven for bare de 21245 første løpeordene i romanen, som i sin helhet er 27269 ord lang. Dette har ingen praktisk betydning for kurvens generelle form, og kurven for romanen i sin helhet viser tilsvarende konveksitet.



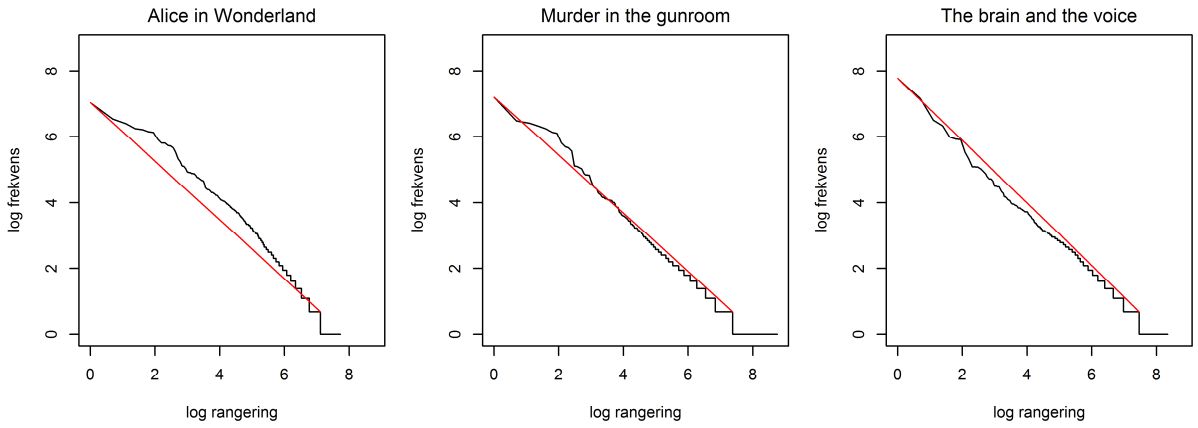
Figur 10-2: Zipf-kurver for lemmafrekvenser i hånd- og tastetekster. Begge aksene er logaritmiske. Ulikhetene når kurvene nærmer seg y = 0, skyldes at tastekorpuset er lengre enn håndkorpuset.

Figur 10-3 viser Zipf-kurver for tre tekster av ulike teksttyper, klippet til lik lengde, 21 245 ord: barneromanen *Alice in Wonderland*[21], voksenromanen *Murder in the gunroom*[22] og den faglitterære voksenboka *The brain and the voice*[23]. De tre diagrammene illustrerer at de tre ulike teksttypene resulterer i tre ulike former på kurvene. Barneromanen er konveks, voksenromanen er lineær, mens fagboka er konkav. De tre formene svarer til en økning fra venstre til høyre i de tre teksttypenes variasjon og spesifisitet i ordforrådet.

[21] *Alice in Wonderland,* teksten hentet fra R-programpakken languageR (Baayen, 2008), som er et tillegg til boka av Baayen. http://cran.r-project.org/web/packages/languageR/index.html

[22] *Murder in the gunroom*, http://www.gutenberg.org/cache/epub/17866/pg17866.txt

[23] *The brain and the voice,* http://www.gutenberg.org/cache/epub/13111/pg13111.txt



Figur 10-3: Ordform-baserte Zipf-kurver for tre engelskspråklige tekster: en skjønnlitterær barnebok, en skjønnlitterær voksenbok og en faglitterær voksenbok. De røde linjene er ikke regresjonslinjer, men linjen fra det mest frekvente ordet til det siste ordet med 2 forekomster. Kurvene viser tre forskjellige hovedtendenser: konveks, lineær og konkav. Alle kurvene viser dessuten uregelmessigheter for de mest frekvente ordene, altså hovedsakelig funksjonsord. Figuren viser også at kriminalromanen har langt flere unike ord enn de to andre, og at de aller mest frekvente funksjonsordene er mer frekvente i fagboka.

Zipfs teorem gjelder tekster av adskillig større lengde enn tekstene i elevtekstkorpuset, og teoremet gjelder dessuten enkelttekster og ikke tekstsamlinger. Det er ikke gitt at leksikalske egenskaper av denne typen gjelder for samlinger av korte tekster på samme måte som for lange enkelttekster. Tvert imot virker det intuitivt sannsynlig at en samling av korte tekster om samme emne har mindre variasjon og spesifisitet i ordforrådet enn én lengre tekst. Elevtekstkorpuset må derfor sammenlignes med andre korpus av samme type før man kan si noe mer om i hvilken grad Zipf-grafens kurve og helning karakteriserer tekstsamlingen på noen måte.

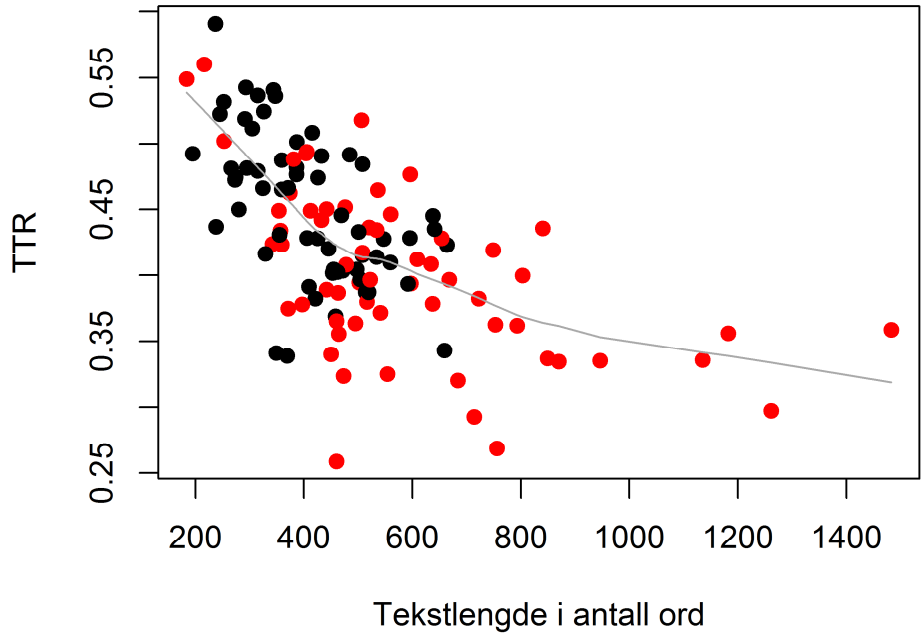
10.2 TTR

TTR er et akronym for *type/token ratio*, som vil si forholdstallet mellom antall typer og antall eksemplarer. I dette kapitlet (10) dreier diskusjonen seg når ikke annet er sagt, om ordformer, altså er TTR i dette tilfellet antall unike ord delt på antall løpeord; definisjonen av ord er basert på CG3. I 10.4.5 drøfter jeg bruken av andre enheter enn de grafiske ordformer i denne sammenhengen.

TTR eller varianter av TTR har vært mye brukt som mål på leksikalsk variasjon eller *diversity*, men i sin grunnleggende form er TTR et ubrukelig mål på variasjon, ettersom det korrelerer systematisk (negativt) med tekstlengde. Holmes og Forsyth (1995, s. 115) finner en korrelasjonskoeffisient på –0,7, og en lignende sammenheng finner vi i elevtekstkorpuset. Sammenhengen er en matematisk nødvendighet, og det trengs bare et øyeblikks refleksjon for å innse det. I svært korte tekster er det bare forskjellige ord; det er altså like mange typer som eksemplarer, og TTR = 1. Etter hvert som teksten blir lengre, vil man begynne å gjenta ord som er brukt tidligere, i første omgang særlig funksjonsord. Etter hvert som teksten vokser seg lengre, blir det stadig vanskeligere å ta i bruk ord som ikke forekommer tidligere i teksten, og forskjellen mellom antall typer og antall eksemplarer øker. I svært lange tekster kan man tenke seg at skribenten etter hvert bruker opp hele sitt ordforråd, eller i hvert fall

den delen av det som er relevant for tekstens emne, og dermed kommer til et punkt der ingen nye ordtyper tilføres teksten. Ingen av tekstene i elevtekstkorpuset er imidlertid av en slik lengde.

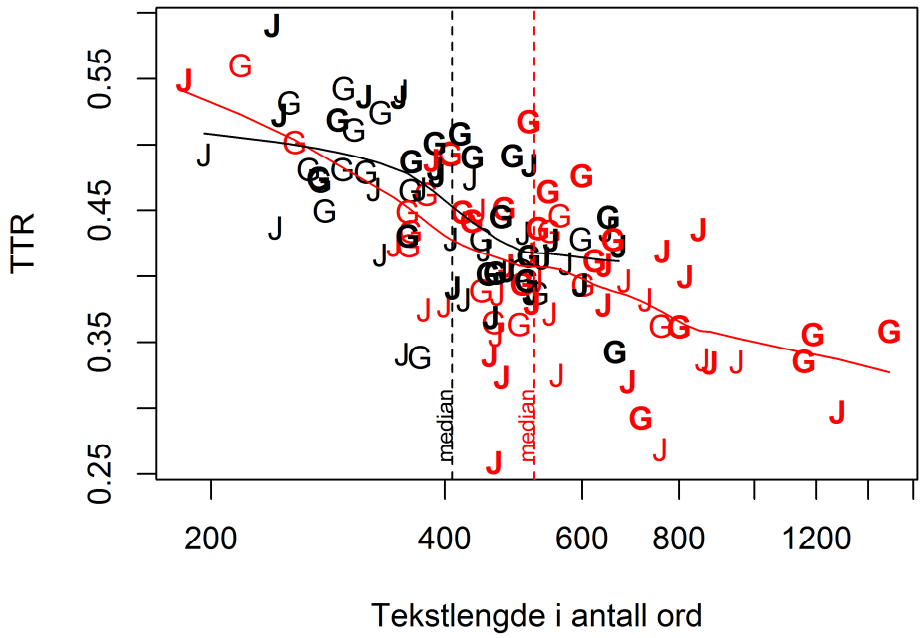
Siden elevtekstlengdene ikke er normalfordelt, beregner jeg Pearsons korrelasjonskoeffisienter mellom TTR og logaritmetransformerte tekstlengder (se 7.2.2.5 om logaritmetransformering) og finner da $R \approx -0{,}66$.[24] TTR og logaritmetransformerte tekstlengder er normalfordelte, ifølge Shapiro-Wilks normalitetstest, $W \approx 0{,}996$, $p \approx 0{,}98$; $W \approx 0{,}988$, $p \approx 0{,}36$.



Figur 10-4: Sammenheng mellom TTR og tekstlengde. Svarte punkter representerer håndtekster og røde punkter tastetekster.

Figur 10-4 viser den klare tendensen til negativ korrelasjon i elevtekstkorpuset mellom TTR og tekstenes lengde målt i antall ord, men det framgår også klart at forbindelsen ikke er lineær, noe som også er intuitivt åpenbart. Hvis den hadde vært lineær, ville det ha medført at regresjonslinjen ville krysse x-aksen, og dermed negativ TTR for tekster som overskrider en viss lengde. Figur 10-5, der x-aksen er tegnet logaritmisk, antyder at korrelasjonen kan ligge nærmere en log-lineær sammenheng. (Se en mer utfyllende diskusjon av dette i 10.3.1.)

[24] Testen er utført på 120 tekster som er skrevet av bare 60 elever. Observasjonene er dermed ikke uavhengige, så anvendt som hypotesetest er testen ugyldig. Jeg oppgir derfor ikke p-verdi.



Figur 10-5: Korrelasjon mellom TTR og tekstlengde på logaritmisk x-akse.

Den sterke negative korrelasjonen medfører altså at TTR på tekster av ulik lengde er en verdiløs variabel, slik Holmes og Forsyth (1995, s. 115) også er inne på. Det finnes flere alternative mulige løsninger på dette problemet.

I en undersøkelse som dette, der én paret faktor er av størst interesse, kan man beskjære den lengste teksten i hvert par og dermed sammenligne TTR i hvert tekstpar der begge tekstene nå er av samme lengde. Man kan dermed sammenligne forskjellene i TTR for hvert par og vurdere om endringene går i samme retning, men forskjellene vil ha usammenlignbare størrelser, noe som medfører at kun en rangeringsbasert analyse er aktuelt. Verdiene og variasjonen, og dermed differansene, vil være størst for korte tekster. Jeg kjenner ikke til noen arbeider som har benyttet eller diskutert dette grepet, og jeg drøfter ikke dette som en relevant løsning i denne avhandlingen.

Et mer drastisk grep er å beskjære alle tekstene til samme lengde n og beregne TTR for de beskårne tekstene (Biber, 1988, s. 238-239). Den korteste teksten i elevtekstkorpuset er 184 ord lang. Det betyr at man må velge $n \leq 184$ for å beholde hele utvalget; eventuelt kan man velge en høyere verdi for n og forkaste fra utvalget alle elever som ikke har skrevet to tekster på minst n ord. Dette reduserer selvfølgelig utvalgets størrelse og kan introdusere skjevheter i det som i utgangspunktet var et balansert utvalg. Biber (1988, s. 238-239) bruker $n = 400$ som minstemål. Denne grensen ville i så fall utelukke 29 av elevene i elevtekstkorpuset, noe som særlig går ut over gruppen av gutter med middels skriveferdigheter, jf. tabell 10-2 nedenfor.

Tabell 10-2: Fordeling av elever på kjønn og skriveferdighet. Til venstre hele korpuset med et balansert utvalg av 60 elever. Til høyre bare elever med to minst 400 ord lange tekster, totalt 31 elever.

	Alle elever			Bare elever med to tekster over 400 ord		
	Middels	Sterk	**Sum**	Middels	Sterk	**Sum**
Gutter	15	15	**30**	4	10	**14**
Jenter	15	15	**30**	8	9	**17**
Sum	**30**	**30**	**60**	**12**	**19**	**31**

En annen type grep er å dele opp alle tekstene i flere segmenter av samme lengde, for eksempel 100 ord, og regne ut TTR for alle segmentene og deretter en gjennomsnittlig segment-TTR for hver tekst (se 10.4.2). Fordelen med denne løsningen er at alle elevene blir med i analysen, men variabelens egenskaper endres fra å være et globalt mål for variasjon til et mer lokalt basert mål for variasjon, og konsekvensene av dette er ukjente.

En tilnærming som opprettholder variablenes globale egenskaper, er å korrigere TTR-verdiene matematisk. Dersom TTR virkelig korrelerer loglineært med tekstlengde, kan TTR-verdien justeres med logaritmen av tekstlengden multiplisert med negativen av stigningstallet til den loglineære regresjonslinjen. Det kan eventuelt finnes alternative matematiske transformasjonsmåter for TTR-verdiene som for eksempel Brunets W (se 10.3.3 nedenfor) eller Hultman & Westmans OVIX (se 10.3.2).

De neste avsnittene evaluerer og sammenligner de ulike alternativene, men før det vil jeg peke på at også konseptuelt er TTR som mål på variasjon problematisk. Intuitivt virker forholdstallet mellom typer og eksemplarer som et naturlig mål for variasjon, men TTR måler egentlig bare antall forskjellige ord i en tekst eller et tekstsegment av en gitt lengde. TTR fanger ingenting av frekvensfordelingen mellom de ordtypene som finnes i teksten eller tekstsegmentet. En TTR-verdi på 0,5 i et tekstsegment på n = 100 forteller at det er 50 ulike ord i segmentet, men TTR-verdien er uberørt av hvorvidt 49 av disse ordene er brukt én gang hver, mens det siste ordet er brukt 51 ganger, eller om alle ordene er brukt 2 ganger hver. Vår intuitive forståelse av variasjon eller kompleksitet er knyttet til også denne type forskjeller, og det er lett å overfortolke TTR-verdiene til å gjenspeile variasjon på flere nivåer enn de faktisk gjør. (Se mer utdypende diskusjon av dette poenget i 10.6.)

10.3 Transformert TTR

Jeg skal først presentere og evaluere noen transformasjonelle metoder for å motvirke sammenhengen mellom TTR og tekstlengde.

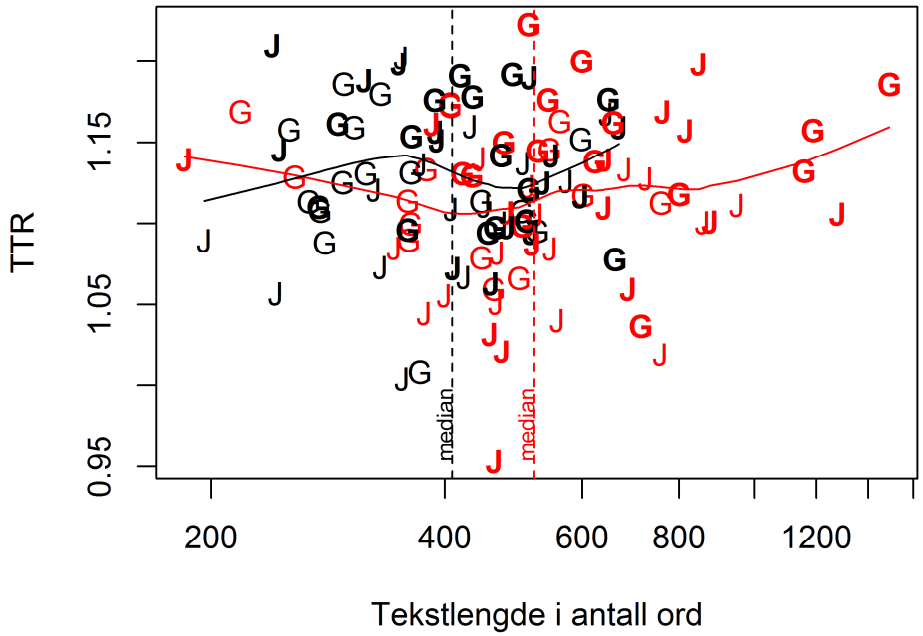
10.3.1 Loglineært korrigert TTR (*log-TTR*)

Figur 10-5 antyder en loglineær sammenheng mellom TTR og tekstlengde. Hvis dette er riktig, kan TTR justeres med produktet av log-tekstlengde og negativen av stigningstallet for regresjonen mellom TTR og log-tekstlengde, noe som skulle både *rette ut* og *flate ut* regresjonskurven i figur 10-4 på side 173. Formelen i (111) nedenfor regner ut estimatet for

stigningstallet i regresjonen, mens formelen i (112) justerer TTR-verdiene etter tekstlengde og estimatet for stigningstallet.

```
(111) TTR.stigning <- lm(lex$TTR~log(lex$n.ordeks))$coefficients[2]
(112) lex$TTR.korrigert <- lex$TTR - log(lex$n.ordeks)*TTR.stigning
```

Den *logaritmekorrigerte TTR* er normalfordelt ifølge Shapiro-Wilks normalitetstest, $W \approx 0{,}982$, $p \approx 0{,}12$. Det er ingen korrelasjon med tekstlengde, slik figur 10-6 viser; Pearsons korrelasjonstest gir $R \approx 0$, som ventet ut fra hensikten med og metoden for korrigeringen.[25]



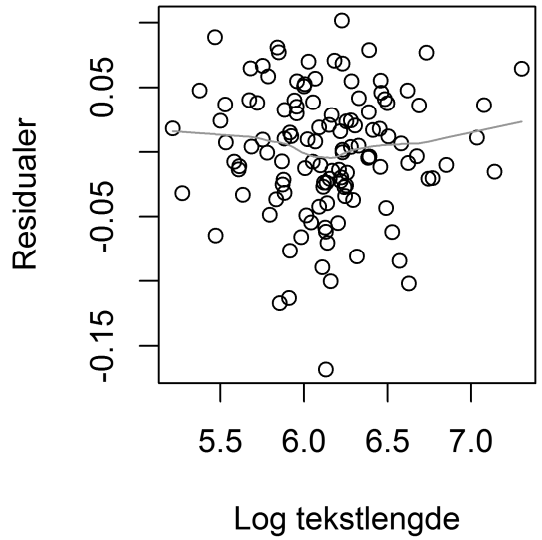
Figur 10-6: Logaritmekorrigert TTR korrelert med tekstlengde på logaritmisk akse

En ulempe med TTR korrigert for tekstlengde på denne måten er at de konkrete verdiene taper konseptuell betydning. At en teksts TTR er 0,4, uttrykker et konkret forhold mellom typer og eksemplarer, selv om man også skal være forsiktig med å overtolke et forholdstall på denne måten. At en teksts *logaritmekorrigerte* TTR er 1,1, har derimot ikke noen naturlig eller lett tolkbar sammenheng med tekstens konkrete egenskaper, og for leser eller forsker skaper ikke verdien noe intuitivt bilde av tekstens karakter. En endring i logaritmekorrigert TTR fra 1,11 til 1,12 virker ubetydelig, men kan være statistisk signifikant og kanskje vesentlig, siden det interkvartile spennet i elevtekstkorpuset bare strekker seg fra 1,09 til 1,15.

Det er også viktig å merke seg at en korrigert modell av denne typen ikke har allmenn gyldighet. Den er utviklet kun for et konkret korpus av tekster, i dette tilfellet skrevet

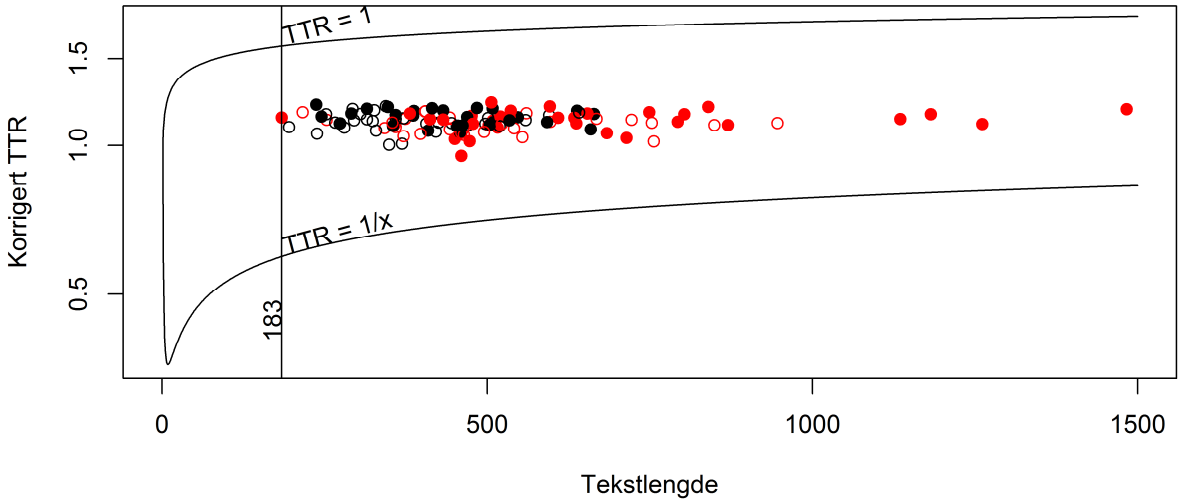
[25] R rapporterer den mer nøyaktige verdien $R \approx 5{,}87*10^{-16}$.

innenfor én sjanger og av et smalt spekter av skribenter. Dessuten er spennet i tekstlengde tross alt ganske lite, fra 184 ord til 1483 ord. Det er ingen grunn til å anta at denne loglineære modellen også vil fungere for betydelig lengre tekster, som romaner, eller for enda kortere tekster.



Figur 10-7: Residualer fra en lineær modellering av sammenhengen mellom logaritmekorrigert TTR og log-tekstlengde.

Dette forbeholdet støttes av figur 10-7, som viser residualene fra den loglineære modellen. Figuren antyder konkavitet, riktignok svak, altså at en (log-)lineær modell ikke passer dataene fullstendig. Dette er også intuitivt rimelig. Heller ikke logaritmeverdiene av TTR kan være lineært synkende med tekstlengde, fordi verdiene også i dette tilfellet vil krysse x-aksen ved tilstrekkelig høye verdier av x. Diagrammet i figur 10-5 illustrerer også at regresjonslinjen vil krysse x-aksen, og ved hjelp av stigningstallet i den lineære modellen kan skjæringspunktet regnes ut til $x \approx 10$, altså en tekstlengde på $e^{10} \approx 22\,000$ ord. En log-lineær modellering av TTR kan dermed ikke være allmenngyldig, selv om den i prinsippet *kan* være tilnærmet gyldig for tekster innenfor et visst spekter av lengder. Figur 10-7 demonstrerer likevel at det er sannsynlig at dataene har en viss, kontinuerlig krumming som også gjelder det aktuelle utvalget, og dette er også intuitivt det mest naturlige.



Figur 10-8: Logaritmekorrigert TTR med teoretiske øvre og nedre grenseverdier. Den vertikale linjen markerer lengden av den korteste teksten i korpuset.

Figur 10-8 viser fordelingen av logaritmekorrigerte TTR-verdier sammen med de teoretiske øvre og nedre grenseverdiene for alle tekstlengder opp til 1500 ord. Den øverste kurven følger de logaritmekorrigerte TTR-verdiene for teoretiske tekster med bare ulike ord, mens den nederste kurven følger de logaritmekorrigerte TTR-verdiene for teoretiske tekster med bare like ord. Diagrammet demonstrerer at det ikke er noen øvre lineær *asymptoteverdi* for variabelen, ettersom logaritmefunksjoner er ubundet. Det er heller ingen nedre asymptoteverdi, men det er en øvre og en nedre teoretisk grenseverdi for enhver tekstlengde, og disse grenseverdiene øker med tekstlengden og er ubundet. Dette viser at det forsøk på tekstlengdeuavhengig linearisering jeg har gjort over, ikke *kan* være allmenngyldig for alle tekster, ettersom verdiene vil krysse den nedre grensekurven når tekstlengden blir stor nok. Figuren demonstrerer også at de logaritmisk korrigerte TTR-verdiene nærmer seg den nedre grensekurven med stigende tekstlengder, noe som kan være en indikasjon på at modellen heller ikke har gyldighet innenfor det aktuelle spekteret av tekstlengder. Modellen har med andre ord både gjort distribusjonen for lineær og for flat; den burde ha justert både konkaviteten og den negative helningen i større grad.

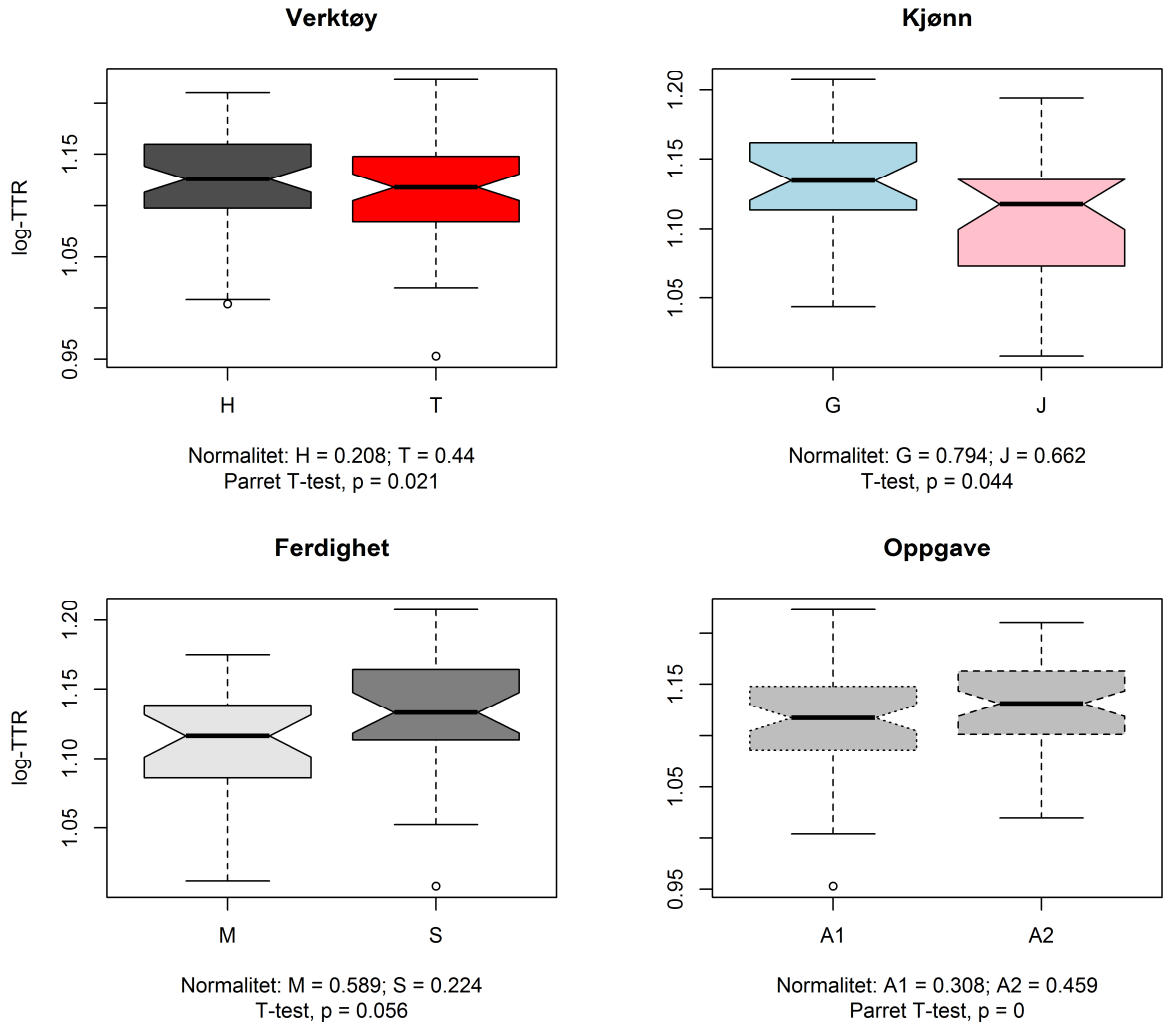
I det følgende presenterer jeg variabelens egenskaper. Jeg bruker i fortsettelsen betegnelsen *log-TTR* for denne formen for loglineært korrigert TTR. Tabell 10-3 nedenfor viser at log-TTR-verdier ligger typisk rundt 1,12, med standardavvik på omtrent 0,05. Utvalget og de relevante delutvalgene er normalfordelt, ifølge Shapiro-Wilks normalitetstest.

Tabell 10-3: Nøkkeltall for loglineært korrigert TTR (log-TTR)

	middel	median	sd	min	maks
Total	**1,117**	**1,118**	**0,049**	**0,945**	**1,220**
Hånd	1,124	1,123	0,045	1,000	1,208
Tast	1,110	1,114	0,051	0,945	1,220
Middels	1,106	1,108	0,043	1,000	1,196
Sterk	1,127	1,134	0,052	0,945	1,220
Gutt	1,128	1,129	0,041	1,006	1,220

	middel	median	sd	min	maks
Jente	1,106	1,106	0,053	0,945	1,208

Figur 10-9 nedenfor illustrerer at variabelen er en del påvirket av alle fire parametre. Håndtekstene har høyere verdier enn tastetekstene, guttene har høyere verdier enn jentene, de sterke har høyere verdier enn de middels elevene, og "Ungdomsfylla"-tekstene har høyere verdier enn "Bøker eller data"-tekstene.



Figur 10-9: Log-TTR fordelt etter fire faktorer

Det er en ganske sterk sammenheng mellom elevenes hånd- og tastetekster, $R \approx 0{,}56$.

En variansanalyse ble utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, antall interaksjonsnivåer begrenset til 2:

(113) `lm(lexD$log.TTR~(kjønn + ferdighet + lengde + forskjell)^2)`

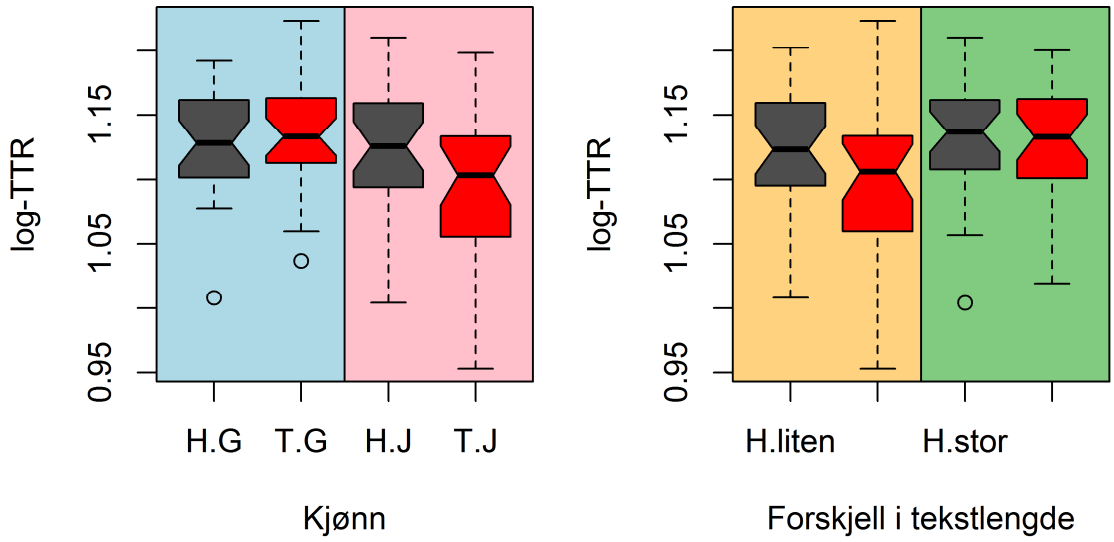
Reduksjon av den maksimale modellen over gav følgende minimale adekvate modell:

```
(114) lm(formula = lexD$log.TTR ~ kjønn + forskjell)

                Df  Sum Sq  Mean Sq F value  Pr(>F)
kjønn            1 0.01691 0.016912   9.682 0.00291 **
forskjell        1 0.00741 0.007407   4.240 0.04405 *
Residuals       57 0.09956 0.001747
---
Multiple R-squared:  0.1963,    Adjusted R-squared:  0.1681
F-statistic: 6.961 on 2 and 57 DF,  p-value: 0.001973
```

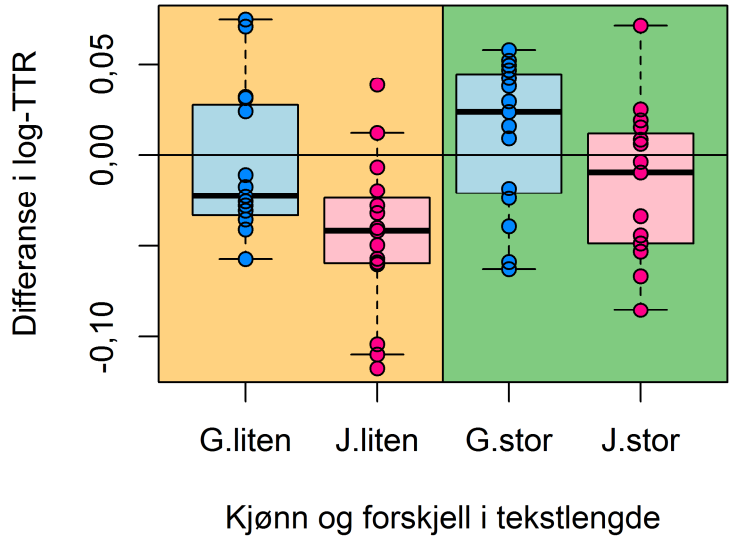
Foruten den generelle tendensen til lavere variabelverdier i tastetekstene, som kom fram i figur 10-9, er det en sterkt signifikant effekt av kjønn og en svakt signifikant effekt av forskjell i tekstlengde. Det er ingen interaksjon mellom de to prediktorene. `Gvlma` (se 7.2.2.4) viser at premissene for anova-analysen er oppfylt (se appendiks A4).

Selv om analysen tyder på at det er en generell effekt av verktøyet, viser boksdiagrammene i figur 10-10 nedenfor at det bare er jenter og elever med liten forskjell i tekstlengde som har lavere verdier i tastetekstene. Gutter og elever med stor forskjell i tekstlengde er upåvirket av verktøyet, og den generelle effekten som analysen gir, skriver seg fra at effekten i disse to segmentene er såpass sterk.



Figur 10-10: Log-TTR. Interaksjon mellom verktøy og kjønn til venstre, og mellom verktøy og forskjell i tekstlengde til høyre.

Siden det ikke er noen interaksjon mellom kjønn og forskjell i tekstlengde, er trolig det nederste segmentet av tastetekstene dominert av jenter som også har liten forskjell i tekstlengde. Figur 10-11 nedenfor viser en viss tendens til en slik effekt, men den er ikke særlig utpreget. Det mest påfallende resultatet er at bare 2 av disse jentene har høyere verdier i tastetekstene enn i håndtekstene, og kanskje at de 3 med de laveste verdiene alle tilhører denne kategorien.



Figur 10-11: Differanse i log-TTR. Interaksjon mellom kjønn og tekstlengdeforskjell. Boks nummer to fra venstre viser at bare to jenter med liten forskjell i tekstlengde har høyere log-TTR i tastetekstene.

Sett i lys av at log-TTR trolig er under-korrigert med hensyn til påvirkningen fra tekstlengde, er det en del usikkerhet knyttet til effektene av tekstlengde som er kommet frem. Også kjønnseffekten må man være kritisk til, siden jentene i utvalget skriver noe lengre tekster enn guttene. Men gitt at analyseresultatene gjenspeiler reelle egenskaper ved tekstene, viser de altså at en del jenter har mindre leksikalsk variasjon i tastetekstene sine enn i håndtekstene, og at det samme gjelder noen av de elevene som *ikke* har særlig stor forskjell i lengde mellom de to tekstene. Det siste delresultatet er muligens litt kontraintuitivt, siden man kanskje kunne vente at det var de elevene som skriver mye lengre på tastatur, som også får mindre variasjon i tastetekstene sine. Men det ser altså ut til at de som skriver mye lengre på tastatur, opprettholder skrivemønsteret sitt for denne variabelen, mens de som skriver omtrent like langt, får mindre variasjon på tastatur enn for hånd. Forskjellen mellom kjønnene er d ≈ 0,80, som er blant de sterkeste effektene i materialet, mens forskjellen relatert til forskjell i tekstlengde er d ≈ 0,50.

Mange forskere har forsøkt å justere TTR-verdiene matematisk på lignende måter som over. I de to neste avsnittene skal jeg se på to av disse forsøkene, nemlig Hultman og Westmans OVIX (Hultman & Westman, 1977) og Brunets W, slik denne er presentert av Holmes og Forsyth (1995) og Baayen (2001).

10.3.2 Hultman og Westmans OVIX

Hultman & Westman (1977, s. 56) skriver at de innen det svenske prosjektet *Skrivsyntax* har utarbeidet et mål på ordvariasjon som er relativt uavhengig av tekstlengde, i hvert fall innenfor de tekstlengder det er snakk om i de svenskstilene og den bruksprosaen de undersøker. Gymnasiasttekstene i materialet deres har lengder mellom 229 og 1309 ord, med middelverdi 588 (Hultman & Westman, 1977, s. 53), altså ganske sammenlignbart med elevtekstkorpuset i denne avhandlingen. De kaller målet for ordvariasjonsindeks, eller OVIX, som er definert gjennom følgende formel (s. 264):

$$V = N^{2-N^{(\frac{1}{OVIX})}}$$

der N representerer antall ordeksempler i teksten og V antall ordtyper. Formelen for OVIX blir dermed:

$$OVIX = \frac{1}{log_N(2 - log_N(V))}$$

I R ser formelen slik ut:

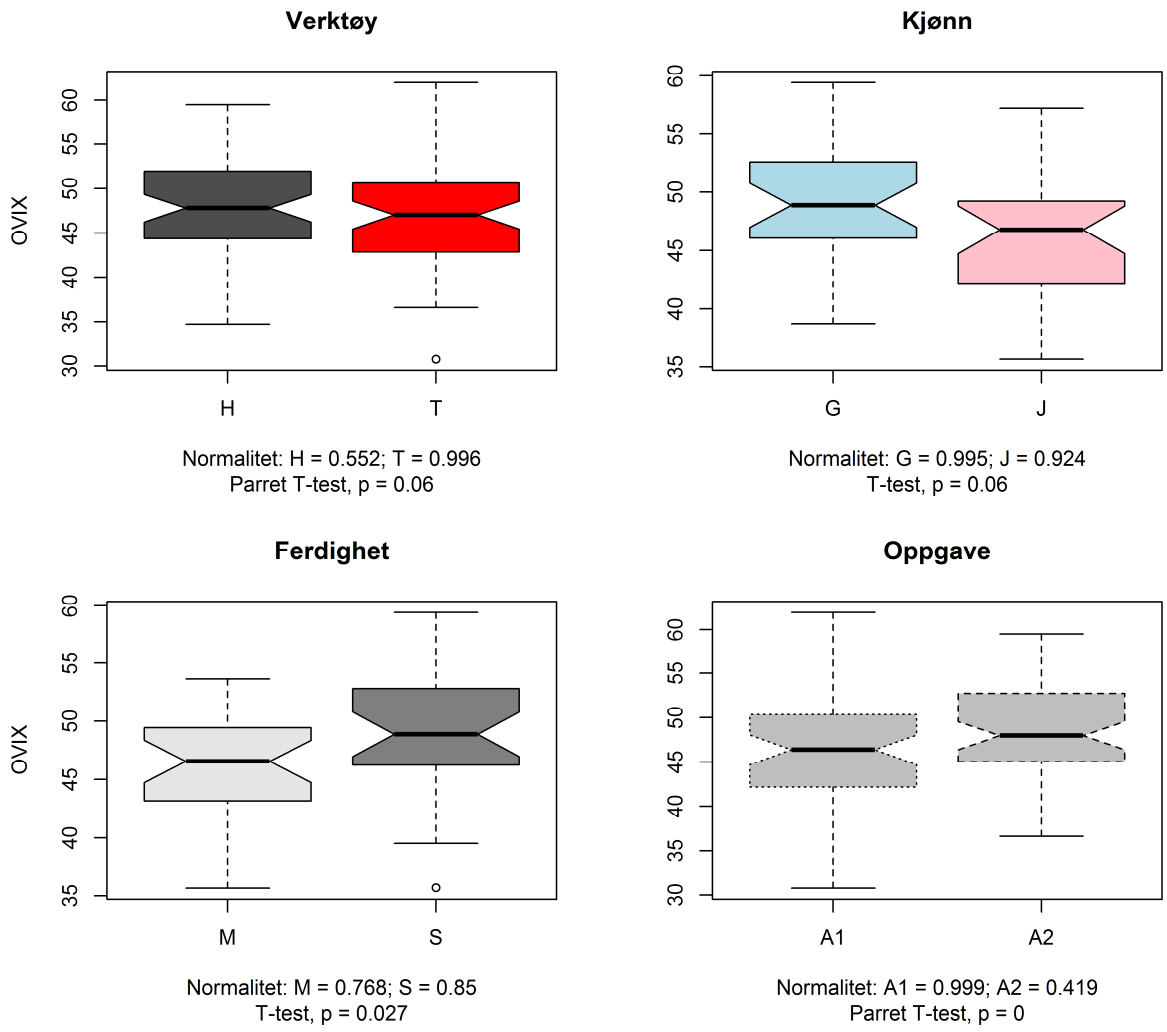
```
(115) OVIX <- 1/log(2-log(V,N),N)
```

Som tabell 10-4 nedenfor viser, ligger median- og middelverdiene for OVIX i elevtekstkorpuset rundt 47, med et standardavvik på i underkant av 6. Alle tallene er basert på ordformer. Utvalget og de relevante delutvalgene er normalfordelt, ifølge Shapiro-Wilks normalitetstest.

Tabell 10-4: Nøkkeltall for OVIX

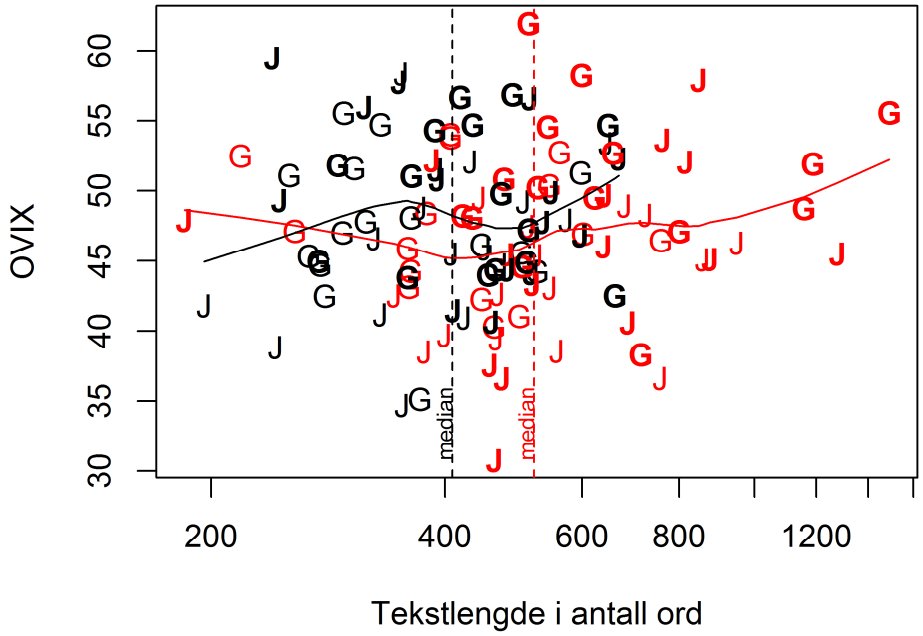
	middel	median	sd	min	maks
Total	**47,3**	**47,1**	**5,8**	**30,3**	**62,0**
Hånd	48,0	47,8	5,6	34,6	59,5
Tast	46,5	46,5	6,0	30,3	62,0
Middels	45,8	45,8	5,1	34,6	57,8
Sterk	48,7	49,2	6,2	30,3	62,0
Gutt	48,5	48,1	5,2	35,1	62,0
Jente	46,1	45,9	6,2	30,3	59,5

Figur 10-12 nedenfor viser svake tendenser til høyere OVIX i håndtekstene og blant guttene. OVIX er imidlertid klart høyere blant de sterke elevene og i "Ungdomsfylla"-tekstene. Alle disse tendensene ligner tendensene for log-TTR.



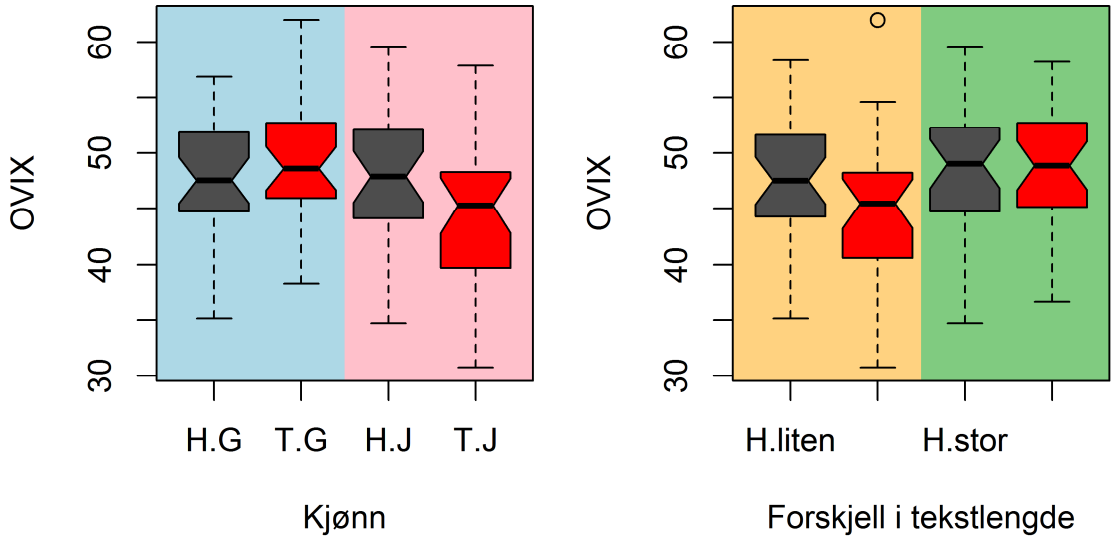
Figur 10-12: OVIX (*Ordvariationsindex*) fordelt etter fire faktorer

Figur 10-13 nedenfor viser at OVIX, som forutsatt for det svenske materialet, også i det norske materialet er relativt uavhengig av tekstlengde, $R \approx 0{,}088$.



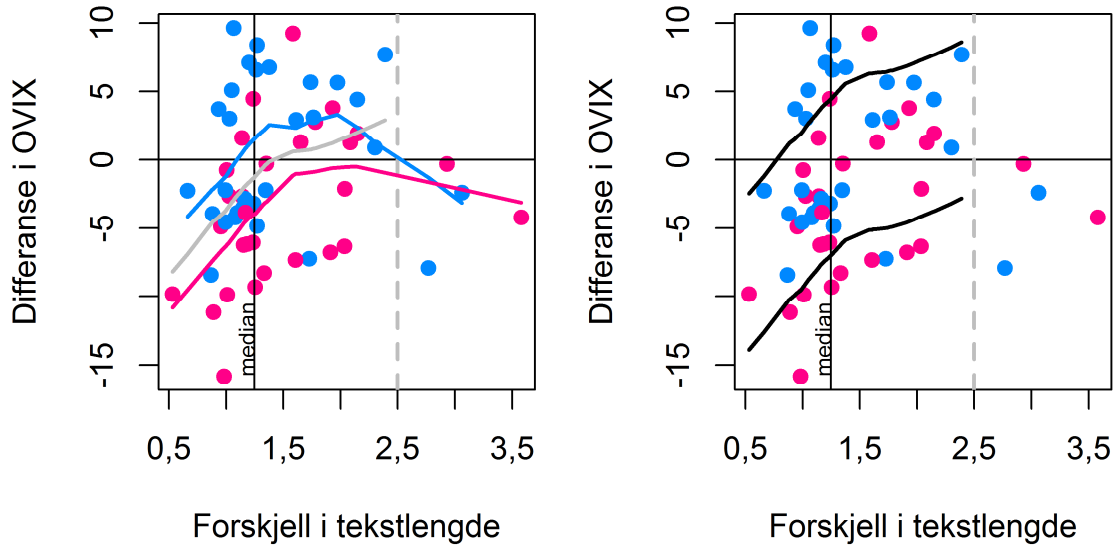
Figur 10-13: OVIX og sammenheng med tekstlengde

Jeg har foretatt en variansanalyse av samme type som for log-TTR med differansen mellom OVIX-verdiene som responsvariabel. Resultatet (gjengitt i appendiks A2) er påfallende likt resultatet for log-TTR (10.3.1), og boksdiagrammene i figur 10-14 nedenfor illustrerer hvor sammenfallende disse variablene er når man sammenligner diagrammene med figur 10-10. Pearsons korrelasjonstest gir da også $R \approx 0{,}992$ mellom de to variablene, noe som tilsier at de to transformasjonene er nærmest kongruente.



Figur 10-14: OVIX. Boksdiagrammer som viser resultatene av anova-modelleringen. Til venstre interaksjonen mellom skriveverktøy og kjønn. Til høyre interaksjonen mellom skriveverktøy og tekstlengdeforskjell. Sammenlign med log-TTR i figur 10-10.

Figur 10-15 nedenfor viser en ganske klar sammenheng mellom forskjell i tekstlengde og differanse i OVIX.



Figur 10-15: Differansen i *OVIX*. De blå og rosa kurvene er regresjonskurver tegnet med `lowess` for gutter og jenter, mens de svarte kurvene markerer regresjonen for hele utvalget pluss/minus ett standardavvik. Den stiplete grå linjen skiller ut fire ekstreme verdier som oppfører seg annerledes enn resten av utvalget.

De kjønnsdifferensierte regresjonskurvene ser ut til å stige gjennom store deler av verdiomfanget for tekstlengdeforskjell, men når en topp og faller ved de største verdiene. Imidlertid kan det se ut til at disse kurvene blir sterkt påvirket av 4 ekstreme verdier for tekstlengdeforskjell. Den grå kurven gjelder for begge kjønn, men bare for verdier av tekstlengdekvotient under 2,5; den er jevnt stigende for hele dette verdiomfanget. Diagrammet til høyre illustrerer også at kjønnsforskjellene kanskje er enda klarere enn variansanalysen viser. De svarte kurvene følger stigningen til regresjonskurven for hele utvalget (unntatt de 4 ekstreme verdiene), men henholdsvis 1 standardavvik over og 1 standardavvik under. Vi ser da at når vi på denne måten kompenserer for korrelasjonen mellom tekstlengdeforskjell og differanse i *OVIX*, er 8 av 9 elever *over* den øverste kurven gutter, mens 8 av 9 elever *under* den nederste kurven er jenter. Også områdene i nærheten av de svarte kurvene har den samme tendensen til skjevfordeling, så dette er en reell egenskap ved utvalget og ikke en artifakt som fremkommer gjennom tilfeldig plassering av de svarte kurvene. Diagrammet illustrerer at kovariansanalyse (*ancova*) antagelig ville være en bedre egnet metode enn anova til å analysere dette materialet. (Se f.eks. Crawley (2007, s. 489-).)

Nyström finner – ikke overraskende – at *OVIX* i gymnasiastmaterialet hennes varierer med sjanger (Nyström, 2000, s. 178). Mer overraskende er det kanskje at også forskjellene mellom kjønnene varierer med sjanger; i noen sjangre har guttene høyest *OVIX*, mens i andre er det jentene som har høyest *OVIX*. I sjangeren "*debattartikel*", som vel er den som kommer nærmest sjangeren i min studie, er det jentene som har høyest *OVIX* med 64, mens guttene har 61, altså motsatt kjønnstendens av hva jeg har funnet. Dessuten er verdiene vesentlig høyere enn i mitt materiale. At verdiene er høyere hos Nyström, har trolig flere årsaker:

elevenes alder[26], tidsrammen, og det faktum at en del av skrivingen stammer fra "*det nationella provet*", noe som trolig påvirker elevens innstilling til arbeidet. Språklige forskjeller mellom norsk og svensk har neppe stor betydning.

Det er altså ikke slik at skriveverktøyet i noen særlig grad har en generell påvirkning på OVIX, men for jenter og for elever som ikke skriver vesentlig lengre på tastatur, synker OVIX i tastetekstene. Når det gjelder påvirkningen fra forskjell i tekstlengde, antyder diagrammene i figur 10-15 ovenfor at sammenhengen kunne framstått enda klarere om jeg ikke hadde dikotomisert denne prediktoren for variansanalysen, men i stedet brukt en kovariansanalyse med forskjell i tekstlengde som en kontinuerlig variabel.

OVIX er et forsøk på å lage en tekstlengdeuavhengig ordvariasjonsvariabel. Dette ser vellykket ut, ettersom det ikke er noen korrelasjon mellom tekstlengde og OVIX i mitt materiale. Imidlertid er det vanskelig ut fra dette materialet å unngå effekten av sammenheng mellom skriveferdighet og tekstlengde. Hvis vi regner med at skriveferdighet og tekstlengde korrelerer, og vi vet at OVIX og skriveferdighet korrelerer, skulle vi regne med en viss korrelasjon mellom tekstlengde og OVIX. Det at vi ikke ser en slik sammenheng, tyder på at OVIX – som log-TTR – underkompenserer for tekstlengde.

Et helt annet moment er at OVIX bare er tenkt å være konstant innenfor et visst utvalg av sjangre. Vi vet at variabler som har med ordvariasjon å gjøre, varierer sterkt med sjanger. Jeg er på bakgrunn av dette sterkt tvilende til at det går an å finne et teoretisk tekstlengdenøytralt ordvariasjonsmål.

10.3.3 Brunets W

Holmes og Forsyth (1995, s. 115) refererer til Brunet (1978)[27] som hevder at *W* er en tekstlengdeuavhengig og forfatterspesifikk indeks, der

$$W = N^{V^{-a}}$$

I formelen står *N* for antall eksemplarer og *V* for antall typer, som i formlene for OVIX i forrige avsnitt, mens *a* er en konstant mellom 0,165 og 0,172. Holmes og Forsyth brukte $a = 0{,}170$ i sin studie, og Baayen (2001, s. 26) bruker det samme.

Det virker usannsynlig at Brunets *W* skulle være uavhengig av *N* også for svært korte tekster. Dersom tekstene er så korte at $V = N$, er *W* i hvert fall ikke konstant, så spørsmålet er heller

[26] Jeg har faktisk ikke vært i stand til å finne eksplisitte opplysninger om elevenes alder i Nyströms undersøkelse, utover at de alle går i gymnasieskolan. Det er mulig at dette indikerer at elevene er fordelt utover de tre årstrinnene, eller at de alle befinner seg på det øverste årstrinn, men sannsynligvis er deres gjennomsnittsalder høyere enn gjennomsnittsalderen på elevene i min studie.

[27] Brunet, E. (1978). Vocabulaire de Jean Giraudoux: Structure et Evolution. Slatkine, Geneve.

for hvilken minste *N* *W*-kurven flater ut. Dette kan sjekkes ved å gjøre en korrelasjonstest mellom *W* og økende *N* i en tekst, slik jeg er gjort for log-TTR i diskusjonen på side 195. Et diagram som plotter *W* mot segmenter med økende *N* av samme tekst vil også kunne avdekke eventuelle uheldige sammenhenger. Baayen (2001, s. 27) har gjort nettopp dette for *Alice in Wonderland* og avdekker at *W* øker med *N* (se også Baayen, 2008, s. 224). I elevtekstkorpuset korrelerer også *W* ganske sterkt med *N* for alle Brunets verdier av *a*, R mellom 0,52 og 0,59 med Pearsons korrelasjon.

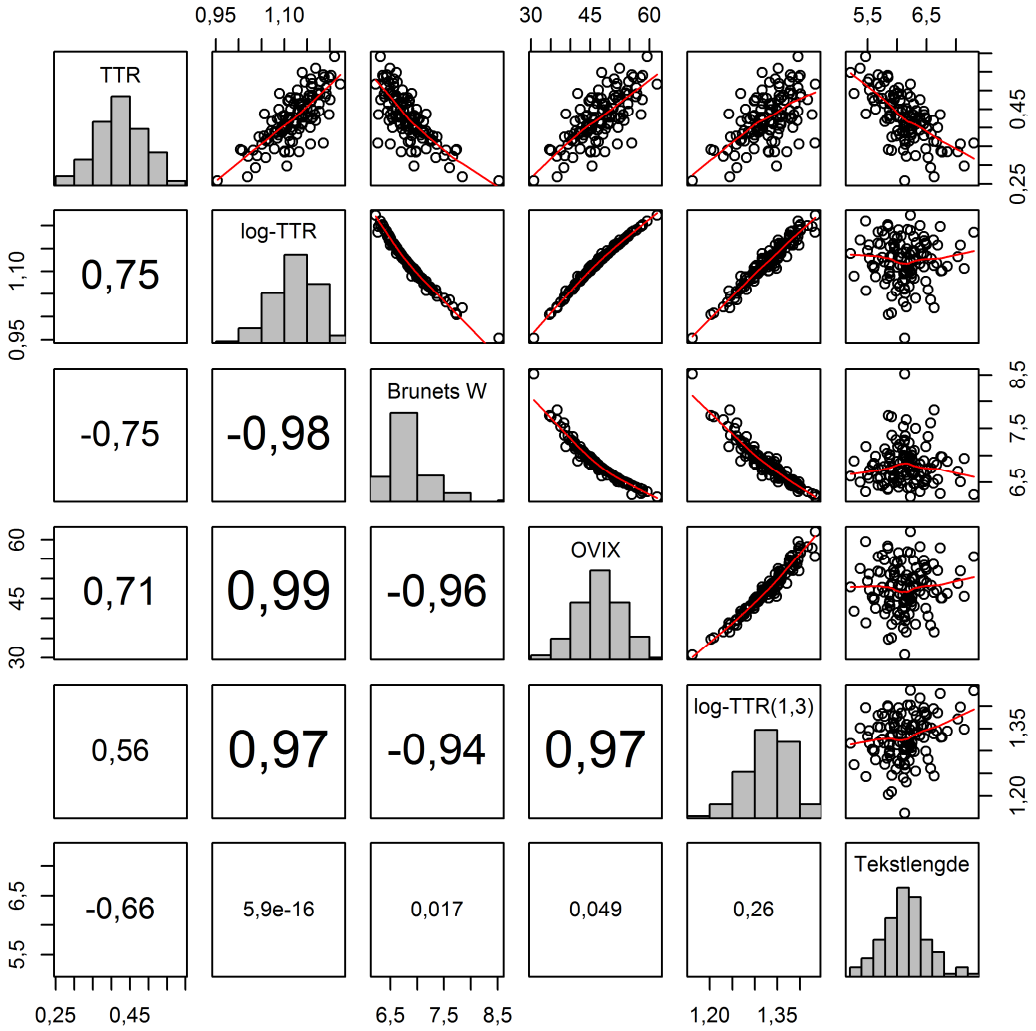
Baayen (2001, s. 26) kritiserte Brunets *W* blant annet på grunnlag av at konstanten *a* ikke har noen "*sensible interpretation*", men at den bare er en heuristisk verdi som er valgt fordi den gir de ønskede egenskapene. I og med at *W* øker med tekstlengde, er det rimelig å anta at Brunet valgte en *a* som passet med de tekstlengdene han studerte, som var vesentlig større enn tekstlengdene i elevtekstkorpuset. Hvis Brunets *W* skulle kunne brukes i andre tekstkorpus, måtte man derfor velge en *a* som ville vise seg å ha de ønskede egenskapene for de tekstlengder det da ville være snakk om. Ved å la *a* variere også utenfor de verdiene som Brunet nevnte som aktuelle, har jeg funnet at $a = 0{,}22$ gir liten korrelasjon med tekstlengde i elevtekstkorpuset. I likhet med *OVIX* samsvarer også denne justerte varianten av Brunets *W* svært tett med *log-TTR*, og alle analyseresultatene (se appendiks A3) ligger tett opptil både *log-TTR* og *OVIX*, selv om de konkrete variabelverdiene selvfølgelig er forskjellige. Dessuten er Brunets *W* speilvendt i forhold til *log-TTR* og *OVIX*, slik at høyere verdier representerer *mindre* leksikalsk variasjon. Tabell 10-5 nedenfor viser at middelverdiene ligger i underkant av 7, med standardavvik noe under 0,4. Alle tallene er basert på ordformer. Variabelen er høyreskjev og ikke normalfordelt, så selv om differanseverdiene som er brukt i anova-analysen er normalfordelte, bør resultatene av analysen brukes med en viss varsomhet.

Tabell 10-5: Nøkkeltall for Brunets *W* med $a = 0{,}22$

	middel	median	sd	min	maks
Total	**6,85**	**6,80**	**0,37**	**6,23**	**8,60**
Hånd	6,79	6,77	0,30	6,32	7,75
Tast	6,91	6,85	0,42	6,23	8,60
Middels	6,92	6,88	0,32	6,38	7,83
Sterk	6,79	6,70	0,41	6,23	8,60
Gutt	6,77	6,74	0,29	6,23	7,72
Jente	6,94	6,90	0,42	6,28	8,60

10.3.4 Oppsummering

De tre metodene for matematisk justering av *TTR* er i realiteten svært nært beslektede transformasjoner. Korrelasjonstester (se figur 10-16 nedenfor) viser korrelasjonskoeffisienter vesentlig høyere enn 0,95 mellom alle tre, og analyseresultatene fra anova-modelleringen er i praksis sammenfallende. Ingen viser systematisk sammenheng med tekstlengde i elevtekstkorpuset, nettopp slik forutsetningen for alle metodene var.



Figur 10-16: Diagram som viser sammenhenger mellom ulike matematisk justerte TTR-baserte variasjonsmål: TTR, log-TTR, Brunets W, OVIX, justert log-TTR (se 10.3.5 nedenfor) og logaritmen av tekstlengde. Tallene er Pearsons korrelasjonskoeffisienter. Diagrammet viser den tette sammenhengen mellom log-TTR, Brunets W og OVIX, og hvordan alle disse har null-korrelasjon med tekstlengde i elevtekstkorpuset. Alle variabler har ikke-normal distribusjon; dessuten består utvalgene av 120 parvist avhengige observasjoner, så de nøyaktige R-verdiene må også tolkes i lys av dette.

Tendensen for de tre variablene er at de utviser lavere verdier i tastetestene for jenter og for elever som skriver mye lengre på tastatur. Punktdiagrammer antyder at både kjønnstendensene og de lengderelaterte tendensene kanskje ville kommet enda klarere fram dersom jeg i stedet for anova hadde benyttet en kovarians-analyse med tekstlengdekvotient som kontinuerlig prediktor.

Det virker altså som jentene som gruppe skriver med mer repetisjon når de skriver på tastatur. Den tallmessige forskjellen skriver seg her ikke bare fra at et utvalg av jentene har en avvikende språkbruk fra resten av utvalget, for når man tar lengdeparameteren med i betraktning, viser det seg at det dessuten er slik at guttene dominerer den øvre delen av

diagrammet. Men guttene har altså tilnærmet like variasjonsverdier i hånd- og tastetekster, mens utvalget av jenter som helhet får lavere verdier i tastetekstene.

Det samme gjelder for elever som skriver omtrent like langt med begge verktøy. Disse elevene skriver med mindre variasjon i tastetekstene, mens de som skriver lengre på tastatur, beholder det samme skrivemønsteret med begge verktøy. Dette virker i utgangspunktet noe overraskende, ettersom hypotesen er at raskere produksjon resulterer i mer spontant og dermed mindre variert språk. Jeg har tenkt på to forklaringer på dette med litt ulike men relaterte perspektiv.

Som nevnt i 8.4.2 skriver de fleste elevene lengre på tastatur. En mulig årsak til at tasteteksten ikke blir lengre hos enkelte av elevene, kan være at eleven ikke "har noe å skrive om", altså mangler momenter eller motivasjon. Manglende motivasjon kan føre til mer uinspirert skriving og mindre oppfinnsom språkbruk, mens få momenter naturlig vil føre til mer repetisjon av leksikalske ord.

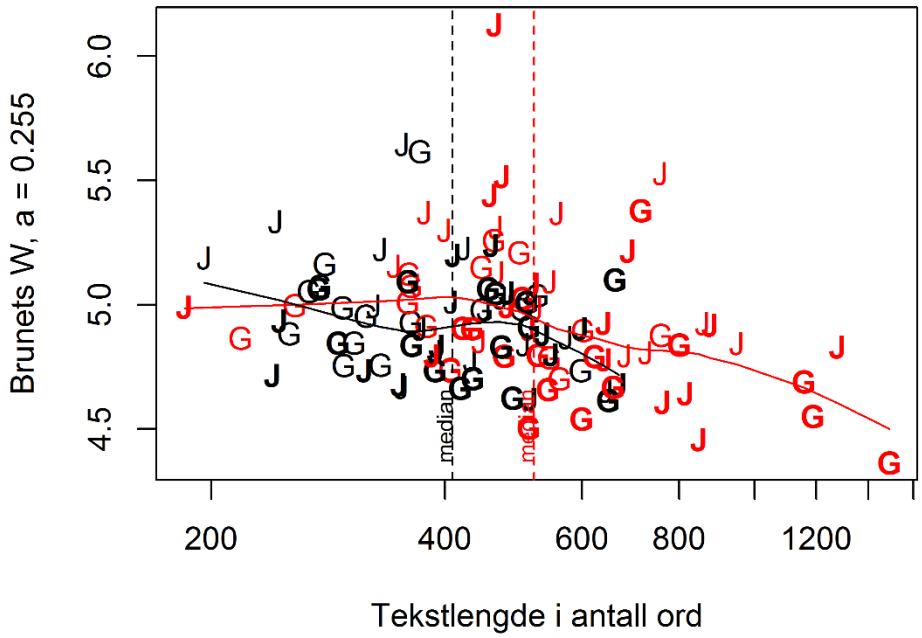
Jeg regner med at alle eller så godt som alle elevene skriver raskere på tastatur. 54 av elevene svarer "viktig" eller "litt viktig" på spørsmålet i spørreskjemaet om høyere hastighet er en grunn til at de foretrekker pc som skriveverktøy; ingen benekter at høyere hastighet er en realitet, men 4 av elevene svarer ikke på spørsmålet. Man kan tenke seg at de som skriver mye lengre på tastatur, også er de som skriver raskest på tastaturet og dessuten på andre måter behersker tekstbehandlingsverktøyet best, og at disse elevene dermed får utnyttet de mulighetene til å utvikle teksten som ligger i det.

10.3.5 Diskusjon

Fremstillingen over viser at mye tyder på at transformasjonene ikke i tilstrekkelig grad kompenserer for den negative sammenhengen mellom TTR og tekstlengde, og kanskje heller ikke for konkaviteten i TTR. Dette gjelder alle de tre variantene av transformert TTR, og det innebærer at det vil være en tendens til at lange tekster har fått verdier som gjenspeiler mindre variasjon enn de i realiteten inneholder. Dette har trolig ikke særlig stor innvirkning på resultatet når det gjelder kjønn. Selv om jenter skriver noe lengre enn gutter, gjelder dette først og fremst i håndtekstene, og forskjellen i påvirkning på differansevariabelen vil dermed være ganske liten.[28] Når det gjelder prediktoren tekstlengdeforskjell, kan det imidlertid tenkes at konsekvensene er større. Differansen for elever med liten tekstlengdeforskjell kan ha blitt kunstig forsterket, mens differansen for elever med større tekstlengdeforskjell kan ha blitt kunstig redusert. I så fall vil dette kunne ha hatt konsekvenser for utfallet av anova-modelleringen, og hele diskusjonen omkring fortolkning av tekstlengdeeffekt vil i så fall være irrelevant.

[28] Siden håndtekstene generelt er kortere enn tastetekstene, er feilmarginen mindre for disse tekstene.

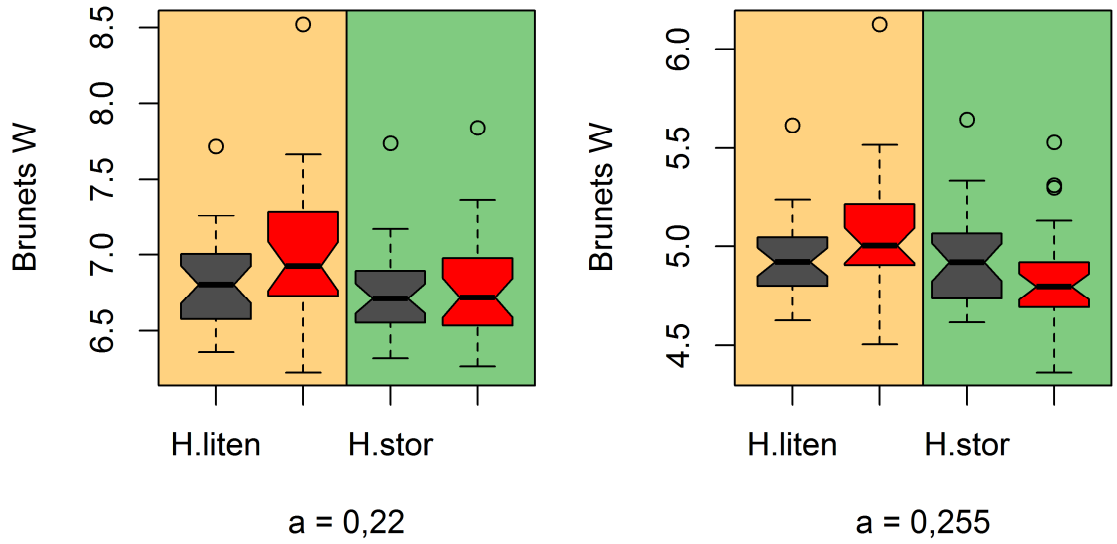
En måte å eksperimentere med dette på, er å justere verdien av *a* i Brunets *W*. Ved å velge *a* = 0,255 oppnår man en moderat korrelasjon med tekstlengde, som vist i figur 10-17 nedenfor.



Figur 10-17: Brunets *W* for *a* = 0,255. Moderat korrelasjon med tekstlengde, R ≈ –0,31. Lavere *W* representerer større variasjon.

Brunets *W* er speilet i forhold til *log-TTR* og *OVIX*, slik at lavere *W*-verdier representerer mer variasjon. Diagrammet viser dermed en positiv korrelasjon mellom variasjon og tekstlengde, R ≈ 0,31 (rho ≈ 0,29), og *a* = 0,255 er valgt slik at korrelasjonen blir omtrent av samme styrke som for det teoretisk lengdenøytrale $FSTTR_{W=400}$ (rho ≈ 0,31, se 10.4.1 nedenfor).

En sammenligning av boksdiagrammer for *a* = 0,22 og *a* = 0,255 i figur 10-18 viser nettopp at påvirkningen fra tekstlengdeforskjell er endret, særlig for elever med stor tekstlengdeforskjell.



Figur 10-18: Brunets *W*. Resultatet av anova-modellering med to forskjellige verdier av *a*.

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, antall interaksjonsnivåer begrenset til 2:

(116)
```
lm(lexD$BrunetsW.a0255 ~ (kjønn+ferdighet+lengde+forskjell)^2)
```

Anova-reduksjon resulterer i en minimal adekvat modell der tekstlengdeforskjell spiller en viktigere rolle enn for a = 0,22 ($F \approx 12{,}65$, $p < 0{,}001$):

(117)
```
lm(formula = lexD$BrunetsW.a0255 ~ kjønn + forskjell)

             Df Sum Sq Mean Sq F value   Pr(>F)
kjønn         1 0.4800  0.4800   9.496 0.003169 **
forskjell     1 0.7984  0.7984  15.795 0.000201 ***
Residuals    57 2.8813  0.0505
---
Multiple R-squared:  0.3073,    Adjusted R-squared:  0.283
F-statistic: 12.65 on 2 and 57 DF,  p-value: 2.85e-05
```

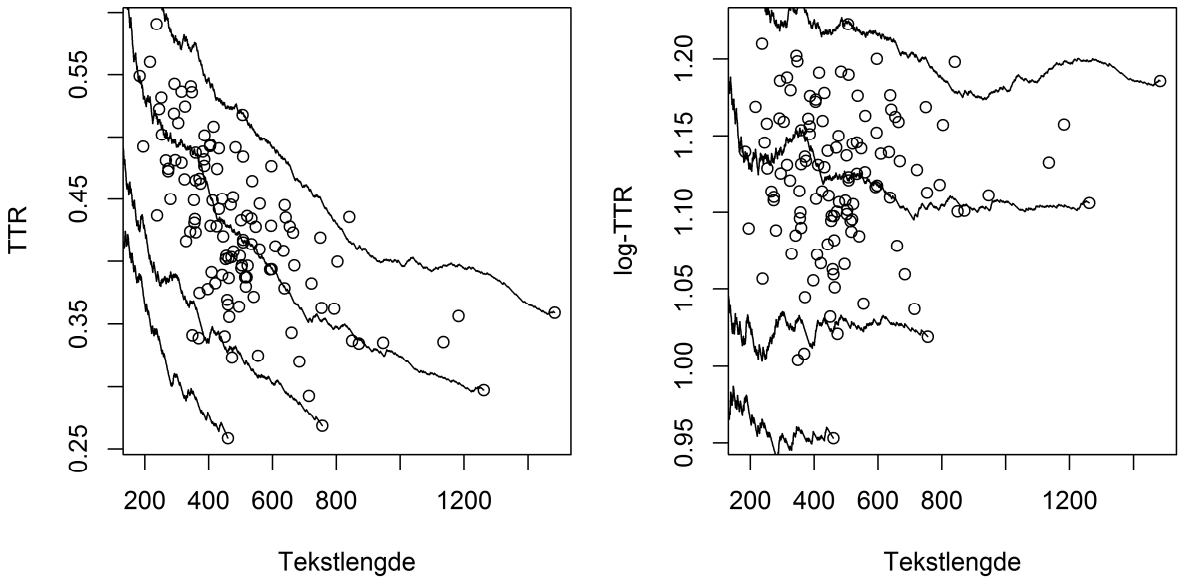
`Gvlma` (se 7.2.2.4) viser at premissene for anova-modelleringen er oppfylt (se appendiks A4).

Problemet med begge disse resultatene for Brunets *W* er at vi strengt tatt ikke vet om *W*-verdiene bør korrelere med tekstlengde, eller eventuelt hvor sterkt de bør korrelere. Vi mistenker at det er en sammenheng mellom leksikalsk variasjon og tekstlengde fordi det gjerne er en sammenheng mellom elevers skriveferdigheter eller elevteksters kvalitet og tekstenes lengde, og fordi vi *antar* at økt tekstkvalitet normalt kjennetegnes blant annet av økt leksikalsk variasjon. Dessuten tyder figur 10-8 (side 178) på at det er et misforhold mellom de faktiske transformerte *log-TTR*-verdiene og de teoretiske grenseverdiene.

En måte å evaluere tekstvariablers gyldighet på med hensyn til uavhengighet fra tekstlengde er å regne ut *progressive verdier* for hver tekst etter hvert som teksten øker i lengde (se for

eksempel Baayen, 2008, s. 224). Under den forutsetning at en tekst har omtrent samme tekstlige egenskap i hele sin utstrekning, bør en de progressive verdiene for en tekstvariabel være ganske stabil for samme tekst i hele det spennet av tekstlengder som variabelen er ment å være stabil for.

For å få et bedre bilde av hvordan TTR kan henge sammen med tekstlengde, kan det være nyttig å beregne og visualisere progressive TTR-verdier for enkelttekster. Dette kan gjøres ved å beregne TTR for de x første løpeordene i teksten, deretter de $x + 1$ første, etc., for alle verdier av x fra 1 til N, der N er antall løpeord i teksten. Progressive TTR-verdier kan slik skape et bilde av hvordan verdier og variasjon utvikler seg etter hvert som teksten blir lengre.



Figur 10-19: Progressive verdier av TTR til venstre og *log-TTR* til høyre. De svarte kurvene følger progressive verdier for fire forskjellige tekster, de samme fire tekstene i de to diagrammene.

Figur 10-19 viser progressive verdier av både TTR og *log-TTR* for fire valgte tekster av ulik lengde og med ulike verdier for *log-TTR*. I diagrammet til venstre demonstrerer de fallende blå kurvene TTR-verdienes naturlig fallende tendens med tekstlengde, mens diagrammet til høyre viser at denne tendensen i stor grad er nøytralisert i *log-TTR*. Man kan likevel se at de to lengste tekstene synes å ha en fallende tendens i de progressive verdiene, i hvert fall fram til $x \approx 800$. Den aller lengste teksten illustrerer dette godt ved at denne tekstens *log-TTR*-verdi er den tiende største i korpuset, mens bare to tekster ligger over kurven for dens progressive verdier, altså over den progressive verdien for de sammenlignbare tekstlengdene av den lengste teksten. Dette støtter antagelsen om at *log-TTR* ikke kompenserer for tekstlengde i tilstrekkelig grad.

Progressive diagrammer kan også brukes til å belyse hva slags tekstegenskaper som påvirker TTR eller andre leksikalske variabler. For den lengste teksten ser vi at både TTR og *log-TTR* stiger bratt rundt $x = 1000$. Stigningen går fra ord 996 til 1024 (tegnsettingen er fjernet):

(118) hadde fått full flidenes panikk ringt heimevernet ringt politiet sluppet hundene løs og sendt tre personlige klagebrev til de høyere makter altså gud jeg virker kanskje en smule streng [A2-210]

I (118) er alle ordene som er unike i hele teksten, understreket, og det går frem at av disse 29 løpeordene er 17 unike i teksten. Jeg synes eksemplet illustrerer at ikke *all* variasjon i *alle* sammenhenger er et symptom på kvalitet. I et kåseri kunne nok denne sekvensen vært underholdende; i et leserbrev kan den raskt gjøre formålet med teksten uklar.

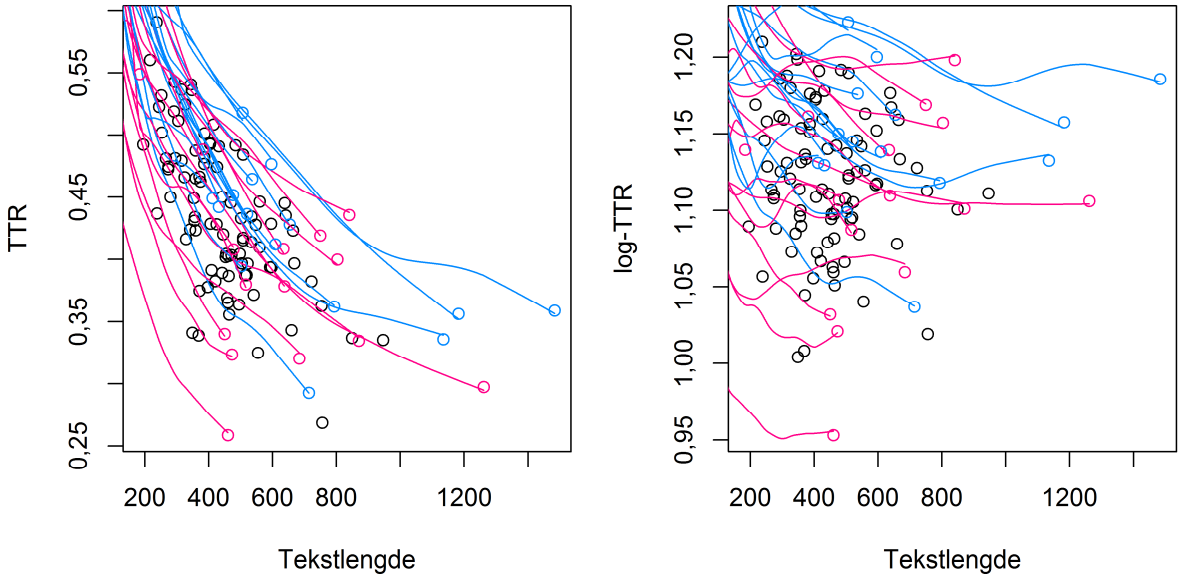
De progressive TTR-verdiene viser også at den teksten som har desidert lavest log-TTR, har svært lave progressive verdier gjennom hele teksten fra $x > 184$. Tabell 10-6 viser toppen av frekvensordlista for denne teksten, både ordformbasert og lemmaformbasert

Tabell 10-6: Frekvenslister for A1-243, teksten med lavest *log-TTR*. Teksten er skrevet av ei jente med karakteren 5 i norsk. Lista viser både lemmaliste og ordformliste.

lemmaform	**antall**	**ordform**	**antall**
være	25	er	24
lese	19	det	18
det	18	ikke	17
data	17	jenter	17
gutt	17	å	17
ikke	17	at	16
jente	17	gutter	16
å	17	som	14
at	16	bøker	12
som	14	i	11
bok	12	jeg	11
jeg	12	men	11
i	11	med	10
men	11	data	9
med	10	leser	9
mye	9	alle	8
si	9	dataen	8
sitte	9	lese	8
på	8	på	8
all	7	enn	7
enn	7	nok	7
nok	7	og	7
og	7	sitter	7
drive	6	jo	6
jo	6	også	6
like	6	driver	5
også	6	foran	5

Lista preges av to egenskaper. For det første er det svært få leksikalske ord i den, bare \lese\, \data\, \gutt\, \jente\, \bok\, \sitte\ og \drive\. Dette er akkurat de samme leksikalske ordene

som finnes i sitatet i "Bøker eller data"-oppgaven.[29] Dette tyder på at teksten er ganske momentfattig eller lite nyanserende. For det andre har de mest frekvente leksikalske ordene svært høy frekvens i forhold til de mest frekvente grammatiske ordene. Dette kunne være et uttrykk for høy leksikalsk tetthet i teksten, men en mer naturlig forklaring er nok liten bruk av synonyme uttrykk eller variasjon av perspektiv, altså liten variasjon blant de leksikalske ordene. Teksten har en verdi for leksikalsk tetthet på ca. 0,38, som er noe lavere enn middelverdien for hele korpuset (som er 0,397, se 9.2.3 på side 147), men høyere enn andre kvartil. Høy leksikalsk tetthet er altså ikke forklaringen på de høye frekvensverdiene for de vanligste leksikalske ordene.



Figur 10-20: Progressive TTR-verdier for tastetekster skrevet av sterke elever. Kjønn er markert med fargen på kurvene. I diagrammene er R-funksjonen `lowess` (se 7.4.4) brukt for å glatte ut kurvene, noe som gjør det lettere å se utviklingstendensene. (I diagrammene medfører glattefunksjonen som er brukt på kurvene, at kurvene ikke treffer punktene for alle tekstene.)

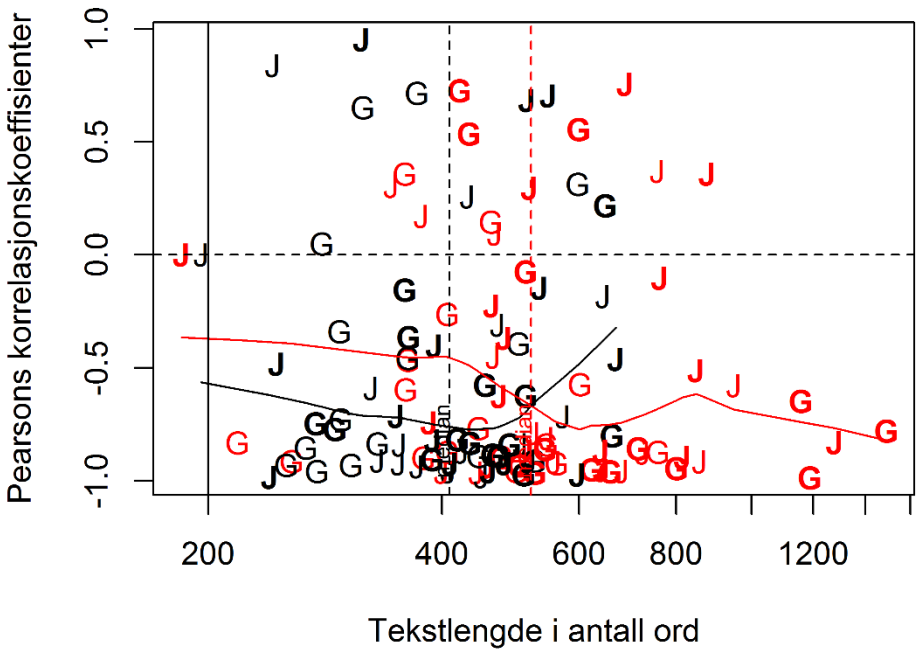
I figur 10-20 er progressive TTR-verdier tegnet inn for et større utvalg av tekstene i korpuset, nemlig for alle tastetekster skrevet av sterke elever. De progressive *log-TTR*-verdiene viser en tydelig fallende tendens; det synes også som om linjene generelt er konkave, altså at de faller mest for lave *x*-verdier. Dette bekrefter at *log-TTR* ikke er korrigert tilstrekkelig hverken for korrelasjon med tekstlengde eller for konkavitet. Kurvene er tegnet for bare et utvalg av tekstene for å gjøre diagrammet mer leselig; tilsvarende tendens synes å fremkomme også for de andre segmentene.

Ved å regne ut vektorer med progressive *log-TTR*-verdier for alle tekstene og beregne korrelasjonskoeffisienten mellom disse vektorene og vektorene av progressive tekstlengder

[29] Scott Jarvis (2015, personlig kommunikasjon) mener det derfor er en fordel at alle ord som forekommer i oppgaveteksten, blir fjernet før analysen. Dette tenkte jeg ikke selv på i en så tidlig fase av prosjektet at det var praktisk mulig å gjøre.

for hver tekst, kan man få et inntrykk av i hvilken grad de progressive verdiene faller med tekstlengde for hele korpuset. Siden *log-TTR* ikke hevdes å være lineær for svært korte tekster, har jeg satt den nedre grensen ved 200 ord, slik at 2 av tekstene ikke er med i beregningen. Den korteste teksten som er med i beregningen, er dermed 216 ord lang, noe som innebærer at korrelasjonskoeffisienten for denne teksten blir beregnet ut fra kun 16 ord.

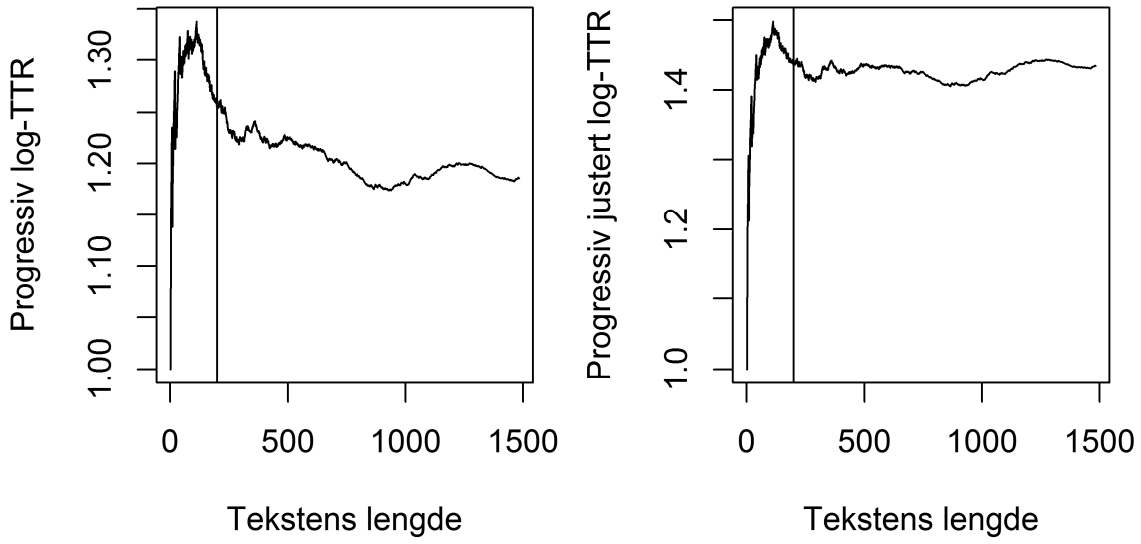
Figur 10-21 nedenfor viser at det er en sterk tendens til at korrelasjonskoeffisientene er negative, og dessuten sterkt negative:



Figur 10-21: Spredningsdiagram over korrelasjonskoeffisienter for progressive *log-TTR*-verdier plottet mot tekstlengde. Diagrammet viser en opphopning av verdier nær –1. De to tekstene helt til venstre er for korte til å være med i beregningen; de er gitt verdien R = 0.

Samtidig er det åpenbart at det ikke gjelder alle tekster; noen har økende *log-TTR* gjennom teksten. Man kunne kanskje ha forventet at korrelasjonskoeffisientene kunne ha blitt påvirket av de store ulikhetene i tekstlengde, men det er ingen klar sammenheng mellom korrelasjonskoeffisienter og tekstlengde.

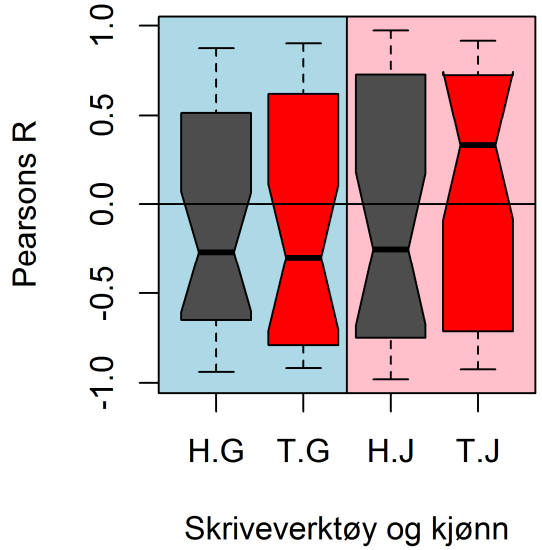
En åpenbar løsning på den uønskede negative korrelasjonen mellom variasjonsmålet og tekstutsnittslengde er å oppjustere stigningstallet i den loglineære transformeringen av TTR, men det er ikke like åpenbart hvor mye stigningstallet bør justeres med. Figur 10-22 nedenfor viser hva som skjer med én av tekstene (A2-210) når stigningstallet for den logaritmiske korrigeringen multipliseres med 1,3; de progressive *log-TTR*-verdiene ser ut til å holde seg stabile etter $x = 200$.



Figur 10-22: Progressive log-TTR-verdier for den lengste teksten i korpuset, tekst A2-210. Til venstre med ujustert stigningstall, til høyre med stigningstallet for den logaritmiske korrigeringen multiplisert med 1,3. Se forklaring i teksten.

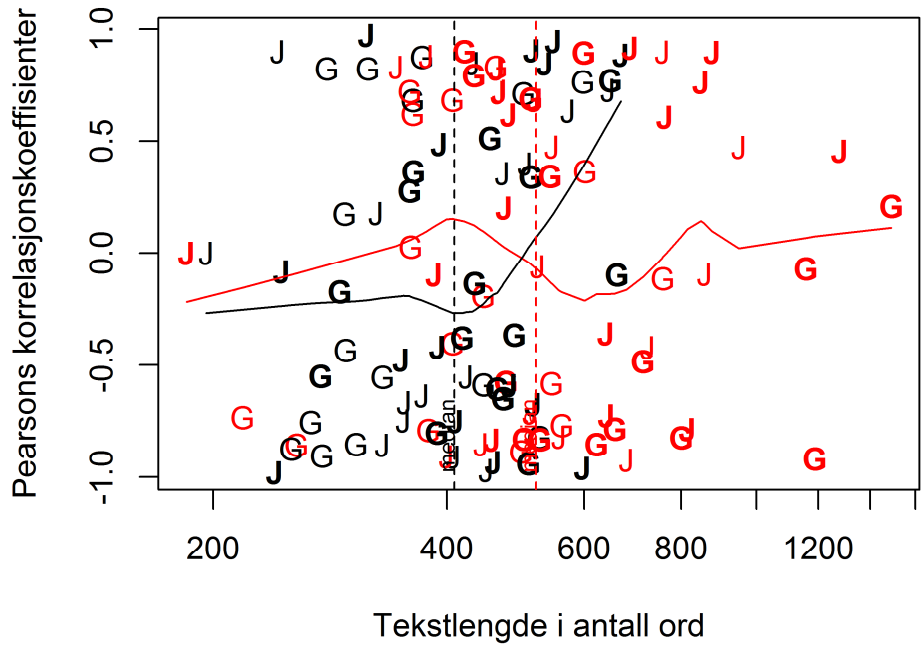
Et boksdiagram over Pearsons korrelasjonskoeffisienter for de progressive justerte TTR-verdiene fra x = 200 og utover og deres sammenheng med en indeks for løpeordene (figur 10-23) viser at den tendensen som gjelder A2-210, gjelder korpuset som helhet, og det er kun små forskjeller når det gjelder verktøy og kjønn, selv om jentenes tastetekster ser ut til å ha noe mer positiv utvikling gjennom teksten enn de andre segmentene. Middelverdien for hele korpuset av 120 tekster er –0,055, og en ettutvalgs t-test gir $t(120) \approx -0{,}88$, $p \approx 0{,}38$, men verdiene er langt fra normalfordelt, så t- og p-verdiene er usikre.[30]

[30] Dessuten er det 120 observasjoner fordelt på 60 elever, så observasjonene er ikke uavhengige. Dette skulle imidlertid ha en reduserende innvirkning på p-verdien, så det er lite tvil om at p-verdien reelt er ganske høy.



Figur 10-23: Korrelasjonskoeffisienter (Pearsons R) for utviklingen av de justerte log-TTR-verdiene fra x = 200 til tekstens slutt. Verdiene er fordelt etter skriveverktøy og kjønn. De to tekstene som er kortere enn 200 ord, har fått tildelt verdien 0.

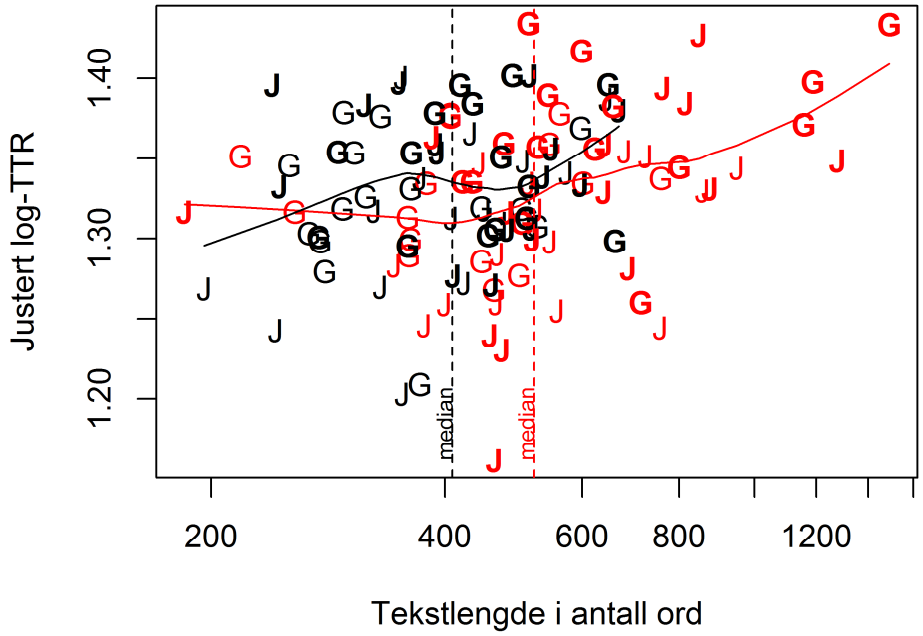
Boksenes utstrekning antyder imidlertid at spredningen er temmelig stor, og et korrelasjonsdiagram mot tekstlengde (figur 10-24) viser sterk bimodalitet; en overvekt av tekstene har R-verdier i nærheten av enten 1 eller –1. Det er med andre ord få tekster som har liten endring i de justerte *log-TTR*-verdiene gjennom teksten.



Figur 10-24: Spredningsdiagram som viser sammenhengene mellom korrelasjonskoeffisienter for utviklingen av justert *log-TTR* fra *x* = 200 til tekstens slutt. De to tekstene som er kortere enn 200 ord, har fått tildelt verdien 0. Diagrammet illustrerer godt R-verdienes bimodalitet.

Faktoren 1,3 ble valgt ganske tilfeldig ved hjelp av visuell inspeksjon av én tekst. Figur 10-23 og figur 10-24 ovenfor viser at justering av stigningstallet med 1,3 trolig er en

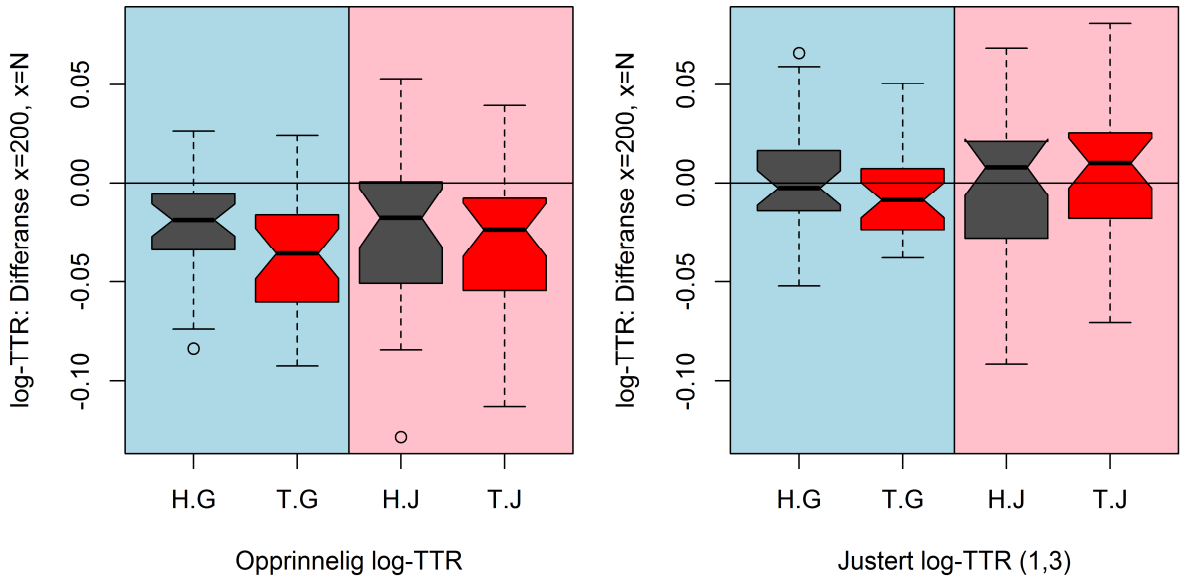
forbedring av *log-TTR*. Figur 10-25 nedenfor viser at den justerte *log-TTR* har en moderat korrelasjon med log-tekstlengde, som forventet og forutsatt, R ≈ 0,26 (rho ≈ 0,24, N = 2 x 60). Denne justeringen retter imidlertid bare opp uforholdsmessigheten i stigningstallet i den lineære regresjonen, mens konkaviteten forblir ujustert. For å rette ut konkaviteten fullstendig måtte trolig logaritmetransformeringen (10.3.1) erstattes av en annen type ikke-lineær transformering. Jeg har ikke gjort forsøk med andre typer transformering enn logaritmisk.



Figur 10-25: Log-TTR justert med faktor 1,3 og sammenhengen med tekstlengde. Pearsons korrelasjonskoeffisient med log tekstlengde gir R ≈ 0,26.

Styrken i korrelasjonen er i området mellom tekstlengdekorrelasjonen for både FSTTR$_{W=400}$ (rho ≈ 0,31) og MOSTTR_LL$_{W=50}$ (rho ≈ 0,22) (se 10.4.1 og 10.4.5.3 nedenfor), som begge skal være matematisk uavhengige av tekstlengde. Det er dermed mye som tyder på at en justering med 1,3 er en forbedring av *log-TTR*, men om dette er den optimale eller mest valide transformeringen av *log-TTR*, er usikkert. Én måte å beregne den beste justeringsfaktoren på er å minimalisere summen av absoluttverdiene eller kvadratene av korrelasjonskoeffisientene, men jeg har ikke gjort denne beregningen. At fordelingen av korrelasjonskoeffisientene for de progressive verdiene er så sterkt bimodal, tyder på at dette målet er for følsomt og ikke særlig godt egnet til formålet, og det er lite poeng i å forsøke å beregne en lite valid variabel mer nøyaktig. En enklere tilnærming er å beregne differansen for hver tekst mellom justerte *log-TTR*-verdier ved $x = 200$ og ved $x = N$. Dersom den justerte *log-TTR* er stabil for ulike tekstlengder, bør denne differansen ligge nær null. Boksdiagrammene i figur 10-26 nedenfor viser at differansen er tydelig lavere enn null (–0,028) for den opprinnelige, ujusterte *log-TTR* (til venstre), mens den er mye nærmere null (–0,000027) når *log-TTR* er justert med faktoren 1,3 (til høyre). Disse differansene er dessuten normalfordelt (W ≈ 0,997, p ≈ 0,99), og en ettutvalgs t-test viser selvfølgelig at avviket fra 0 ikke er signifikant (t ≈ –0,0095, p ≈ 0,899, N = 2 x 60). Både Brunets *W*

(a=0,22) og *OVIX* har den samme egenskapen som den ujusterte *log-TTR*, med verdier som generelt viser større leksikalsk variasjon ved x = 200 enn for tekstenes fulle lengde. (Tallene er ikke vist her.) Dette er akkurat som ventet, gitt det svært nære slektskapet mellom disse to variablene og den opprinnelige *log-TTR*.



Figur 10-26: Boksdiagrammer som viser differansen mellom *log-TTR* ved x = 200 og ved tekstens slutt. Til venstre for ujustert *log-TTR*. Til høyre for *log-TTR* justert med faktor 1,3.

På samme måte som for korrelasjonskoeffisientene kunne man minimalisere avviket fra 0 i et forsøk på å oppnå en mest mulig tekstlengdeuavhengig variabel, men det er fortsatt så stor usikkerhet rundt validiteten i denne beregningen at jeg velger å ikke gjøre dette. For det første er avviket fra nulldifferanse allerede svært lite, og en ytterligere forbedring ville dermed trolig gi relativt liten validitetsmessig gevinst. For det andre er det uklart hvordan de ulike tekstlengdene påvirker differansen, og om differansene burde vektes etter tekstlengde for ytterligere å justere for konkavitet. For det tredje er det risiko for overtilpassing av variabelen om man insisterer på at den skal være fullstendig uavhengig av tekstlengde innenfor et lite tekstkorpus på 120 tekster. Jeg har derfor beholdt faktoren 1,3 som et første grovt anslag og imøteser videre forskning på dette området.

10.3.6 Log-TTR$_{1,3}$ – en justert og forbedret log-TTR

I dette delkapitlet analyserer jeg den justerte *log-TTR*$_{1,3}$ på samme måte som for variablene i foregående delkapitler.

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, interaksjonsnivå begrenset til 2:

(119) `lm(lexD$log.TTR.13~(kjønn+ferdighet+lengde+forskjell)^2)`

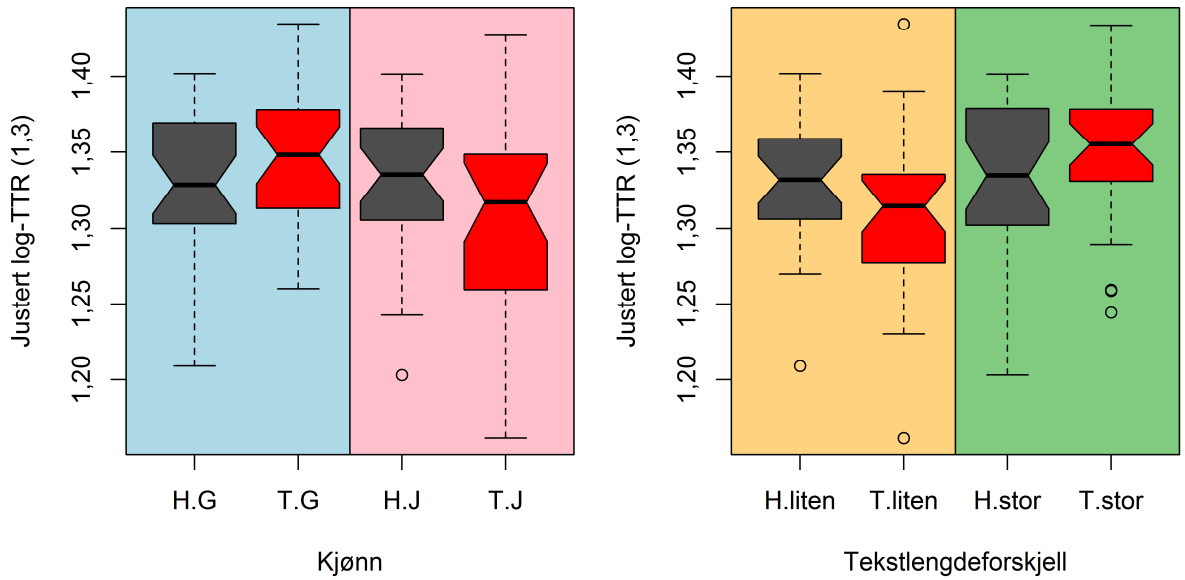
Anova-modellering av den justerte *log-TTR*$_{1,3}$ gir en minimal adekvat modell som er svært lik den for Brunets *W* med *a* justert til 0,255 (se anova-tabellen på side 191):

```
(120) lm(formula = lexD$log.TTR.13 ~ kjønn + forskjell)

              Df  Sum Sq  Mean Sq F value   Pr(>F)
kjønn          1 0.01584 0.015841   8.512 0.005041 **
forskjell      1 0.02648 0.026481  14.229 0.000387 ***
Residuals     57 0.10608 0.001861
---
Multiple R-squared:  0.2852,    Adjusted R-squared:  0.2601
F-statistic: 11.37 on 2 and 57 DF,  p-value: 6.993e-05
```

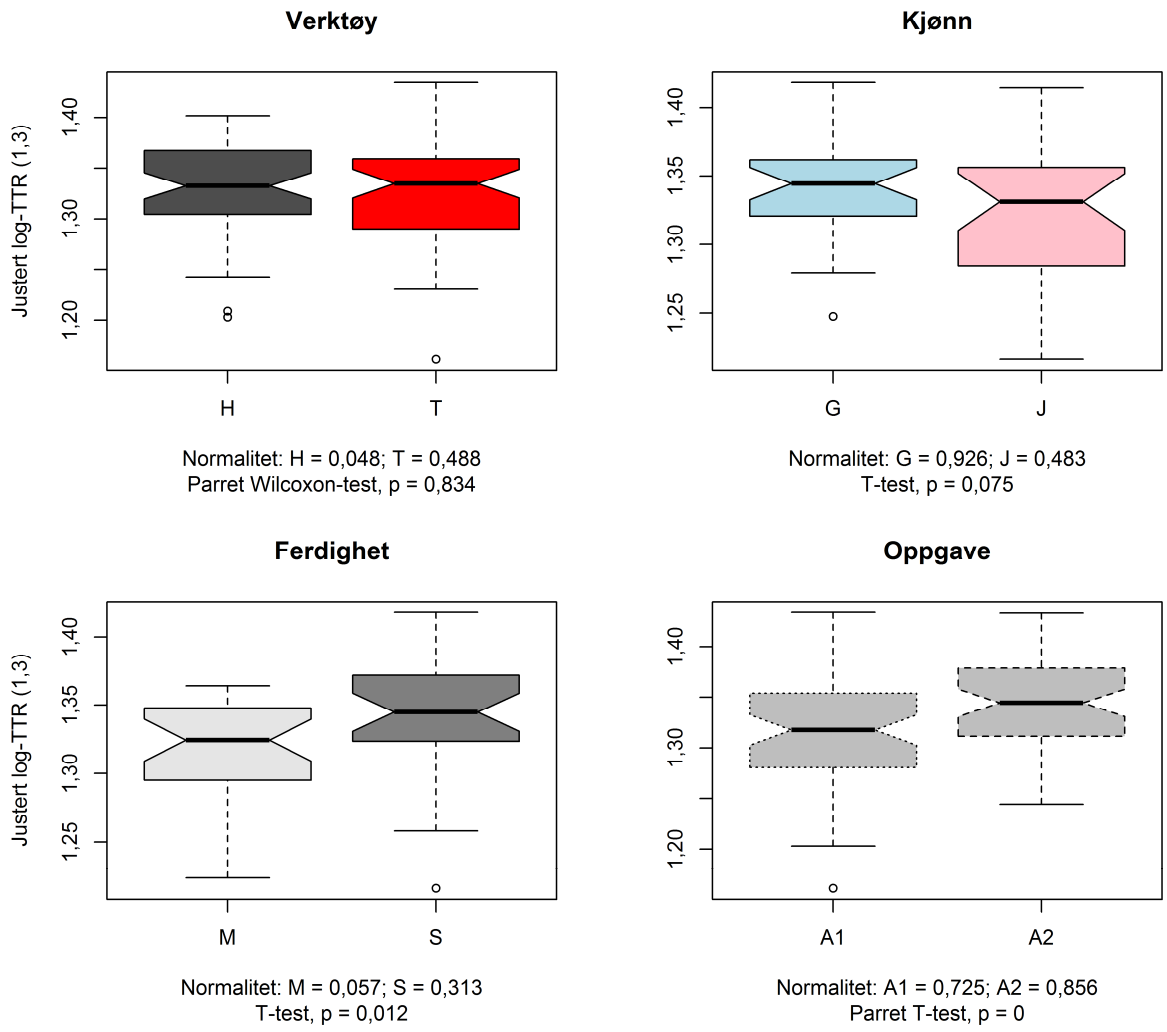
`Gvlma` (se 7.2.2.4) viser at premissene for anova er oppfylt (se appendiks A4).

Effekten av tekstlengdeforskjell er større for *log-TTR*$_{1,3}$ enn for den ujusterte *log-TTR*, og elever med liten tekstlengdeforskjell har lavere leksikalsk variasjon i tastetekstene, mens elever med stor forskjell har høyere variasjon i tastetekstene, som vist i figur 10-27 nedenfor. Når det gjelder effekten av kjønn, har guttene noe høyere leksikalsk variasjon i tastetekstene, mens jentene har lavere leksikalsk variasjon i tastetekstene. Effektene er svært like dem som er presentert for Brunets $W_{a=0,255}$ i figur 10-18 på side 191 ovenfor.



Figur 10-27: Justert *log-TTR*$_{1,3}$ med signifikante resultater fra anova. Til venstre verdier for hånd- og tastetekster fordelt etter kjønn. Til høyre hånd- og tastetekster fordelt etter forskjell i tekstlengde.

Justeringen av *log-TTR* til *log-TTR*$_{1,3}$ fører også til andre generelle egenskaper for variabelen, som vist i figur 10-28 nedenfor. Forskjellen mellom de middels og sterke elevene er nå klarere, mens forskjellen mellom kjønnene ikke lenger er like klar. Det er heller ingen generell forskjell mellom skriveverktøy.



Figur 10-28: *Log-TTR*$_{1,3}$ etter fire faktorer

Totalt vurderer jeg den justerte *log-TTR*$_{1,3}$ som den beste matematisk transformerte TTR-metoden så langt, og det er den jeg benytter i videre analyser og i sammenligninger av ulike variabler, blant annet i prinsipalkomponentanalysen i kapittel 12

Tabell 10-7: Nøkkelverdier for *log-TTR*$_{1,3}$

	middelverdi	median	sd	min	maks
Total	**1,330**	**1,334**	**0,050**	**1,161**	**1,435**
Hånd	1,331	1,333	0,046	1,203	1,402
Tast	1,328	1,335	0,054	1,161	1,435
Middels	1,316	1,318	0,044	1,203	1,401
Sterk	1,344	1,352	0,052	1,161	1,435
Gutt	1,340	1,337	0,044	1,209	1,435
Jente	1,320	1,329	0,054	1,161	1,427

10.4 Segmental TTR

Som vist i forrige delkapittel (10.3) er det gjort mange forsøk på å justere TTR-kurven matematisk for å konstruere en formel for TTR som er uavhengig av tekstlengde. Mye tyder på at dette prinsipielt ikke er mulig, blant annet fordi kurven for lange tekster vil være avhengig av antall momenter og skribentens ordforråd. Når ordforrådet er oppbrukt, vil det ikke tilkomme nye typer, og kurven vil dermed falle raskere enn om skribenten fortsatt produserer nye typer.[31] Dette innebærer at det er teoretisk utelukket at det kan finnes en matematisk formel for allmenngyldig lengdejustering av TTR-kurven. En praktisk, matematisk justering for tekster som ligger innenfor et visst verdiområde av lengder, er det imidlertid mer trolig at er mulig å konstruere for en gitt teksttype. I vårt tilfelle *er* det ganske store forskjeller i tekstlengde, og ettersom tekstlengde også interagerer både med oppgave og med skriveverktøy, som er den parameteren vi ønsker å undersøke, kan en slik justering ha metodisk problematiske konsekvenser, som vi har sett.

I dette delkapitlet presenterer, sammenligner og drøfter jeg ulike metoder for et tekstlengdeuavhengig TTR-mål som *ikke* er basert på matematisk justering av TTR-kurven etter tekstlengde.

10.4.1 FSTTR (Fixed Segment TTR)

Biber (1988, s. 238-239) bruker et enkelt grep for å sammenligne TTR i tekster som ikke er like lange. Han beskjærer alle tekster slik at de blir like lange, før TTR beregnes for de like lange segmentene. Dermed kan TTR-verdiene for ulike tekster sammenlignes direkte. Biber velger 400 ord som grenseverdi, og han må da forkaste fra undersøkelsen sin alle tekster som er kortere enn denne grenseverdien. Biber gir ikke dette TTR-målet noe eget navn, men jeg vil kalle det FSTTR, *Fixed Segment TTR*,[32] for å skille det fra de andre TTR-baserte variasjonsmålene som jeg omtaler i denne avhandlingen.

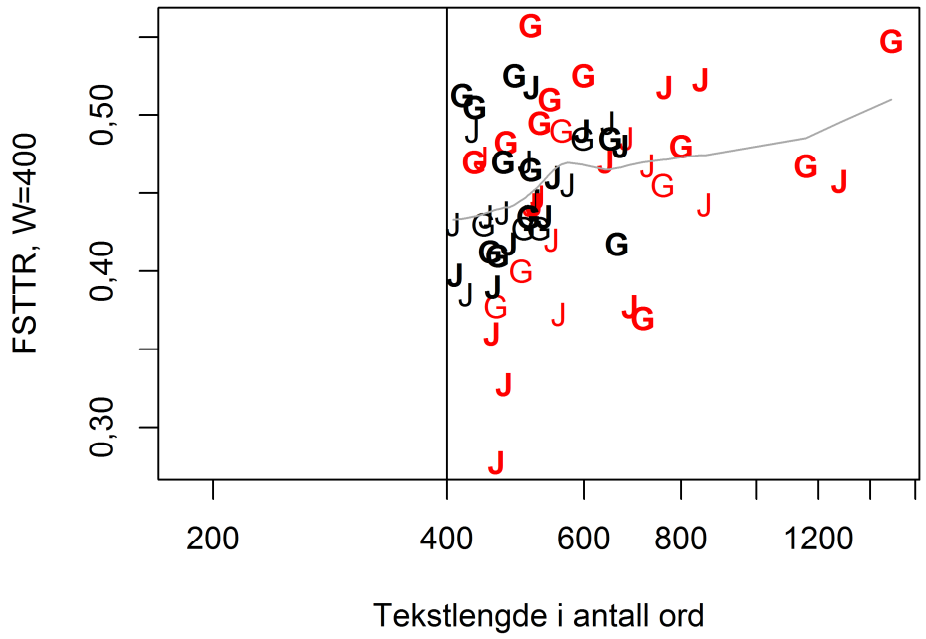
De sentrale fordelene med FSTTR er først og fremst knyttet til to forhold. Det første er at TTR-verdiene ikke blir matematisk påvirket av tekstenes lengde og dermed er umiddelbart sammenlignbare uten matematisk transformering. Det andre er at et segment på 400 ord, i motsetning til i noen av de metodene jeg omtaler under, er langt nok til at TTR-verdiene representerer noe mer enn lokal variasjon. Ulempene ved FSTTR er først og fremst knyttet til at tekster som ikke er lange nok, må forkastes fra undersøkelsen, men man kan heller ikke

[31] For språk med sammensetning som produktiv orddanningsmekanisme, som norsk, er dette ikke fullstendig riktig, siden mengden av ordformer som en språkbruker kan produsere, ikke er prinsipielt avgrenset. I praksis er likevel resonnementet relevant.

[32] Jeg har valgt engelsk terminologi her som en tilpasning til det allerede etablerte MSTTR, for *Mean Segmental TTR*, omtalt under.

se bort fra det at mye av tekstmaterialet må forkastes også fra de tekstene som er lengre enn 400 ord.

Et spredningsdiagram (figur 10-29) viser at den negative korrelasjonen som eksisterer mellom TTR og tekstlengde, ikke gjelder for $FSTTR_{W=400}$.[33]



Figur 10-29: Korrelasjon mellom $FSTTR_{W=400}$ for elevtekstkorpuset og tekstlengde, rho ≈ 0,30, N = 62. Figuren og analysen tar bare hensyn til tekster av elever som har skrevet to tekster som begge er minst 400 ord lange.

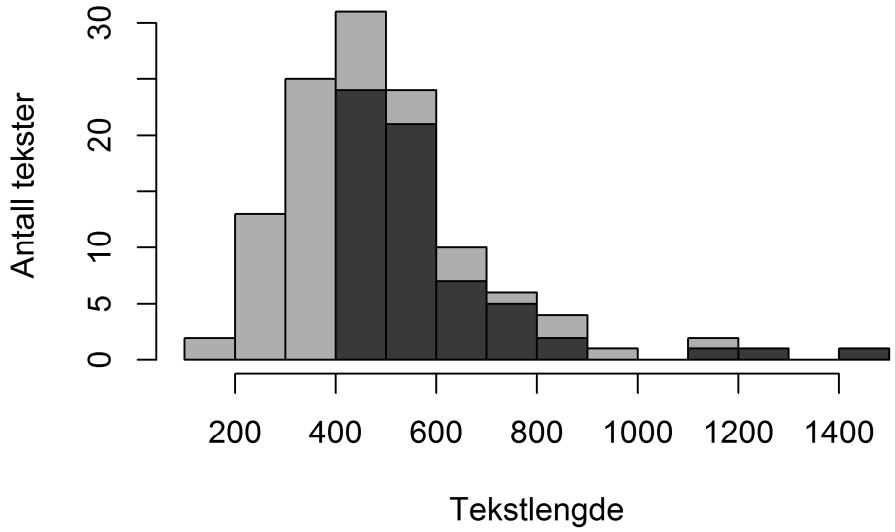
Siden begrensningen av utvalget er gjort på grunnlag av elevenes tekstlengder, er ikke tekstlengde (eller logaritmen av tekstlengde) normalfordelt, slik det går frem av figur 10-30. Jeg har derfor brukt Spearmans korrelasjonsanalyse, og den viser en svak positiv korrelasjon med tekstlengde, rho ≈ 0,30, N = 62. Korrelasjonen mellom tekstlengde og FSTTR skyldes ikke matematiske sammenhenger mellom TTR og antall ord av den typen jeg diskuterte i 10.3, men henger sammen med korrelasjonen mellom tekstlengde og tekstlige egenskaper som kanskje skriver seg fra elevenes skriveferdighet eller tekstens kvalitet eller momentrikdom. Flinkere elever skriver gjerne lengre tekster (8.4.2), og de har også gjerne høyere leksikalsk variasjon, slik det går fram i behandlingen av *log-TTR*$_{1,3}$ i 10.3.6 over, i figur 10-28.

79 av tekstene i elevtekstkorpuset er 400 ord lange eller lengre. Imidlertid forutsetter undersøkelsens problemstilling at analysen foretas på *par* av tekster, så dette krever at alle *elever* som ikke har skrevet to tekster som er minst 400 ord lange, må forkastes. Totalt bare

[33] *W* for *window* brukes gjerne som betegnelse på lengden av et slikt segment av en tekst. Se også 10.4.2. Når ikke annet er spesifisert, bruker alle beregninger for FSTTR i denne avhandlingen grenseverdien W = 400.

31 elever, altså 62 tekster, tilfredsstiller kriteriet. Styrken i analysen blir derfor sterkt redusert, i og med at utvalget blir så godt som halvert.

Det kan være viktig å merke seg at det med denne metoden også blir forkastet mange tekster som er lengre enn 400 ord, slik det går frem av figur 10-30. Diagrammet viser at også temmelig lange tekster er forkastet, slik at mange elever med stor *forskjell* i tekstlengde ikke blir med i analysen.



Figur 10-30: Histogram over tekstlengde. De mørke søylene representerer tekster som inkluderes i undersøkelsen ved bruk av $FSTTR_{W=400}$. De lyse feltene representerer tekster som er forkastet fra utvalget for beregning av FSTTR. Alle tekster under 400 ord er forkastet, men diagrammet viser at også mange tekster over 400 ord er forkastet.

I det hele tatt er denne metoden lite tjenlig for formålet med analysen av elevtekstkorpuset fordi den fører til ikke-tilfeldige frafall fra utvalget. Tabell 10-9 viser at kriteriet gjør skjeve innhugg når det gjelder parametrene kjønn og ferdighet, og dermed svekker balansen i utvalget vesentlig, og tabell 10-8 viser at utslagene som ventet er enda større når det gjelder lengdeparametrene. 23 av 30 elever som generelt skriver kort, er fjernet; elever som skriver kort men har stor forskjell i tekstlengde, er fullstendig fraværende. Derimot er alle langtskrivende elever med liten forskjell i tekstlengde representert.

Tabell 10-8: Utvalg av elever for FSTTR-analyse med grenseverdier W = 300 og W = 400. Tallene viser hvilke skjevheter i lengdeparametrene som introduseres i utvalget ved FSTTR-beskjæring.

	Alle elever			FSTTR, W = 300			FSTTR, W = 400		
	Kort	Lang	**Sum**	Kort	Lang	**Sum**	Kort	Lang	**Sum**
Liten forskjell	18	12	**30**	15	12	**27**	7	12	**19**
Stor forskjell	12	18	**30**	2	17	**19**	0	12	**12**
Sum	**30**	**30**	**60**	**17**	**29**	**46**	**7**	**24**	**31**

Tabell 10-9: Utvalg av elever for FSTTR-analyse med grenseverdier W = 300 og W = 400. Tallene viser hvilke skjevheter i kjønn og skriveferdighet som introduseres i utvalget ved FSTTR-beskjæring

	Alle elever			FSTTR, W = 300			FSTTR, W = 400		
	Middels	Sterke	**Sum**	Middels	Sterke	**Sum**	Middels	Sterke	**Sum**
Gutter	15	15	**30**	8	13	**21**	4	10	**14**
Jenter	15	15	**30**	13	12	**25**	8	9	**17**
Sum	**30**	**30**	**60**	**21**	**25**	**46**	**12**	**19**	**31**

Selv om tekstsegmenter på 400 ord er lange nok til å fange vesentlige egenskaper ved disse elevtekstene, er det dessuten et viktig moment at FSTTR ikke er en global variabel. Leksikalsk variasjon kan være symptom på ulike egenskaper ved teksten, blant annet variasjon i innholdsmomenter. Dersom en tekst på 1200 ord beskjæres til 400 før TTR analyseres, vil denne analysen ikke kunne fange all variasjon som skriver seg fra variasjon i innhold. Dersom den lange teksten for eksempel behandler like mange momenter som en kort tekst, med flere ord per moment, vil FSTTR bare ta hensyn til ordvariasjonen i et mindre antall momenter i den lange teksten. FSTTR vil også gå glipp av avslutningssekvensen i lange tekster, mens avslutningssekvensen vil være med i tekster som er bare litt over 400 ord lange, noe som fører til ulikheter i beregningsgrunnlaget.

For å redusere effekten av at grenseverdien for inkludering resulterer i skjeve utvalg av tekster, kan grenseverdien senkes for eksempel til 300 ord (tabell 10-8 og tabell 10-9). Dette ville øke antall elever i utvalget fra 31 til 46. Imidlertid ville det samtidig resultere i en variabel som forkaster enda mer tekst og i enda mindre grad er global.

På grunn av disse ulempene er FSTTR lite aktuell for analyse i denne avhandlingen. Variabelen er inkludert bare som sammenligningsgrunnlag for de andre TTR-variablene i forbindelse med metodeutvikling. Korrelasjonen mellom tekstlengde og FSTTR er dessuten et viktig premiss for diskusjonen rundt validiteten av de transformerte TTR-målene i kapittel 10.3 ovenfor.

10.4.2 MSTTR (Gjennomsnittlig segmental TTR)

Et intuitivt enkelt grep for å kunne sammenligne TTR også for kortere tekster er MSTTR (*Mean Segmental* TTR), som er omtalt av Malvern, et al. (2004) og trolig ble introdusert av Fairbanks og Johnson i en artikkelsamling i 1944 (Johnson, 1944). Metoden er enkel og nærmest uavhengig av tekstlengde *N*, i hvert fall matematisk. Den består i å dele opp teksten i like lange segmenter eller "vinduer", for eksempel av lengde $W = 100$ ord, regne ut TTR for hvert av de k segmentene og deretter regne ut middelverdien for de k segmentale TTR-verdiene og bruke denne middelverdien som en variabel, kalt MSTTR. Det siste, overskytende segmentet på mindre enn W ord må forkastes.

MSTTR er et langt skritt i retning av et tekstlengdeuavhengig mål for leksikalsk variasjon, og fordelene i forhold til FSTTR (se 10.4.1 ovenfor) er åpenbare. For det første trenger man ikke forkaste noen av tekstene i korpuset, og for det andre er det bare mindre deler av tekstene som ikke blir med i beregningen. Dessuten kan resultater fra ulike undersøkelser sammenlignes, dersom man har valgt samme segmentlengde.

Men det er flere utfordringer også med MSTTR. De viktigste spørsmålene er knyttet til segmentets lengde. Så vidt jeg har kunnet finne ut, er det ingen som har presentert teoretiske begrunnelser for valg av segmentlengden *W*. Fairbanks (1944) og (Wachal & Spreen, 1973) bruker 100, mens Chotlos (referert i (Malvern, et al., 2004)) sammenligner 100, 500 og 1000 og finner sterk korrelasjon mellom resultatene. (Malvern, et al., 2004) bruker 30 og 100 på to ulike teksttyper. (Geizer, 1967) nevner ikke segmentlengde for MSTTR eksplisitt, men han nevner 100 som segmentlengde for et par beslektede metoder. (Covington & McFall, 2010) bruker ikke MSTTR, men et nært beslektet mål som de kaller MATTR (omtalt i 10.4.3 nedenfor), og de anbefaler 500 for stilistiske undersøkelser; imidlertid nevner de at segmentlengder helt ned i 10 og helt opp i 10000 kan være aktuelt til forskjellige formål.

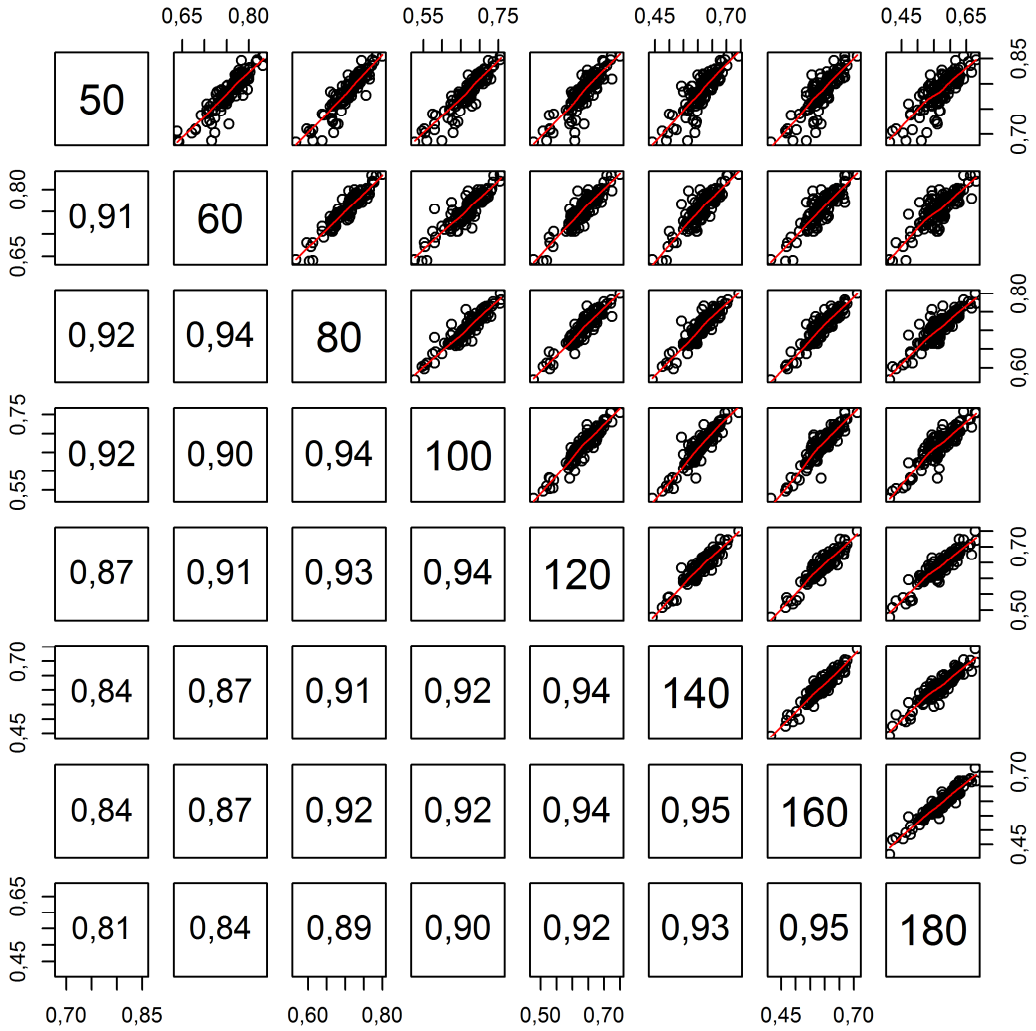
Segmentene må være lange nok til å fange opp variasjon, men de kan ikke være lengre enn den korteste teksten. I denne undersøkelsen må dermed segmentene være ganske korte; av de som nevnes for MSTTR i forrige avsnitt, er bare $W = 30$ og $W = 100$ aktuelle. Med en segmentlengde på $W = 100$ måler man ikke tekstenes globale leksikalske variasjon, men først og fremst lokal variasjon. En tekst som er for eksempel 1000 ord lang, kan ha mye leksikalsk variasjon som skyldes at flere ulike momenter med ulikt leksikalsk tilfang blir behandlet i sekvens gjennom teksten, og slik global variasjon vil fanges opp av for eksempel *log-TTR*, men bare i liten grad av $\text{MSTTR}_{W=100}$, avhengig av hvor lang behandling hvert moment får, sammenlignet med segmentlengden. Momentskifter må falle utenom grenser mellom vinduer for å kunne bidra til MSTTR, og det vil altså være interaksjon mellom vinduslengde og lengden på avsnitt i teksten. I verste fall vil MSTTR ikke i det hele tatt kunne skille mellom en tekst som skriver repeterende om det samme temaet, og en tekst som viser stor tematisk bredde. Dette innebærer at MSTTR ikke måler den samme tekstlige egenskap som et globalt TTR-mål som *log-TTR*. MSTTR måler i større grad en lokal språklig variasjon, mens *log-TTR* i større grad er en global variabel som måler en tekstlig variasjon. Vi kan derfor regne med at lokale, segmentbaserte TTR-mål og globale, logaritmebaserte TTR-mål utfyller hverandre, og begge bør være representert i en analyse av leksikalsk variasjon.

Et annet moment er at det med MSTTR alltid er slutten av teksten som blir forkastet, nærmere bestemt de siste $N - kW$ ordene. Det er åpenbart at et mål som også tar hensyn til slutten av teksten, vil være mer tilfredsstillende, særlig for korte tekster. Dersom slutten inneholder en oppsummering eller konklusjon, vil den kunne ha egenskaper som kan spille en vesentlig rolle for resultatet. Tekster som er kortere enn $2W$, kan få forkastet nesten halve tekstlengden; i elevtekstmaterialet gjelder dette 2 tekster (184 og 195 ord lange) når $W = 100$. For akkurat denne tekstsamlingen er dette et argument for å sette segmentlengden til 90 ord, slik at ingen tekster mister en så stor andel som bortimot halvparten. Den viktigste ulempen med å velge $W = 90$ er at det blir vanskeligere å sammenligne resultatene med andre undersøkelser, ettersom mange undersøkelser bruker W = 100.

Valg av *W* vil dermed ha innvirkning på resultatet på to måter. For det første vil *W* påvirke resultatet på grunn av interaksjon med tekstens lengde. En tekst på for eksempel $N = 190$ ord vil få forkastet nesten halvparten med $W = 100$ men nesten ingenting med $W = 90$. Dersom

tekstens leksikalske variasjon endrer seg fra begynnelsen til slutten, vil dette i praksis kunne ha store konsekvenser for resultatet. For det andre vil ulike *W*-verdier kunne gi ulike MSTTR-verdier på grunn av interaksjoner med tekstenes moment- eller avsnittslengde. Korte momenter kan bidra positivt til MSTTR, mens flere etterfølgende momenter av omtrent lik lengde kan komme til å sammenfalle med segmentgrensene og dermed i liten grad bidra til høyere MSTTR-verdier.

Figur 10-31 nedenfor viser sterke korrelasjoner mellom MSTTR for ulike verdier av *W* mellom 50 og 180, som er den største segmentlengden som er aktuell for elevtekstkorpuset. Korrelasjonen er naturlig nok sterkest for små forskjeller i W, men også for de største forskjellene i *W* er R > 0,8.



Figur 10-31: Korrelasjon mellom MSTTR med ulike *W*-verdier. Diagrammet viser at korrelasjonskoeffisienten R ligger mellom 0,90 og 0,96 for de minste forskjellene i vindusstørrelse, og at korrelasjonen ikke overraskende blir svakere etter hvert som forskjellene i *W* øker. Men selv ganske store forskjeller i *W* gir R-verdier på godt over 0,80. Pearsons korrelasjonstest er brukt, selv om utvalgene alle er noe venstreskjeve og ikke normalfordelt. De 120 observasjonene er heller ikke uavhengige.

Et tredje aspekt, som gjelder både lokale og globale TTR-mål, er at de korrelerer med leksikalsk tetthet; det vil si at TTR påvirkes av forholdet mellom antall leksikalske ord og antall funksjonsord. Høyere andel funksjonsord gir lavere TTR, fordi funksjonsord generelt utgjør små, lukkede klasser og dermed genererer lite variasjon. Det er rimelig å tenke seg at leksikalsk tetthet (kapittel 9 ovenfor) samvarierer sterkere med lokal, språklig leksikalsk variasjon enn med global, tekstlig variasjon, særlig ved små *W*. På den annen side er det begrenset hvor mye repetisjon av funksjonsord som er mulig innenfor smale vinduer som $W = 100$, og korrelasjonen mellom $MSTTR_{W=100}$ og leksikalsk tetthet er bare moderat, rho ≈ 0,28, mens korrelasjonen mellom det globale log-TTR (justert med faktoren 1,3) og leksikalsk tetthet er like stor, rho ≈ 0,33, N = 2 x 60.[34] Dette viser likevel at TTR-baserte verdier måler flere til dels uavhengige tekstlige og språklige egenskaper, til tross for at de gjerne presenteres som et intuitivt lettfattelig mål for leksikalsk variasjon. Eksempler på slike karakteristikker er "a more varied vocabulary for a given length of text" (Biber, 1986, s. 394), "lower type/token ratios, and thus less lexical variety" (Chafe & Danielewicz, 1987, s. 88), "a less varied vocabulary" (Chafe & Danielewicz, 1987, s. 89), "more repetitive in words (lower type/token ratios)" (Horowitz & Berkowitz, 1967, s. 207), "a measure of lexical diversity" (Covington & McFall, 2010, s. 96),

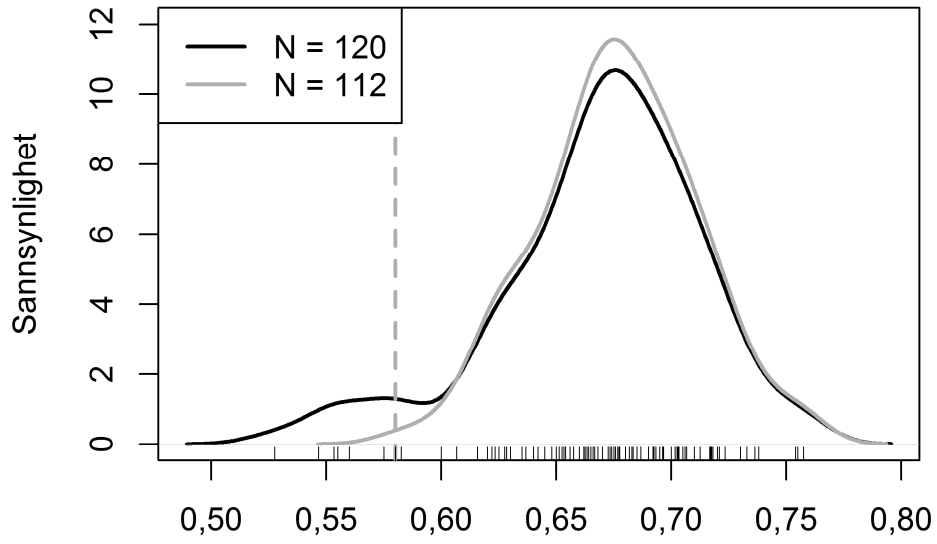
I senere delkapitler skal jeg drøfte måter å møte alle disse tre utfordringene på.

Tabell 10-10: Nøkkelverdier for $MSTTR_{W=100}$

	middelverdi	median	sd	min	maks
Total	**0,668**	**0,675**	**0,044**	**0,528**	**0,758**
Hånd	0,671	0,674	0,043	0,547	0,758
Tast	0,666	0,676	0,046	0,528	0,754
Middels	0,659	0,666	0,040	0,547	0,730
Sterk	0,678	0,682	0,047	0,528	0,758
Gutt	0,681	0,680	0,035	0,547	0,758
Jente	0,656	0,662	0,049	0,528	0,755

Tabell 10-10 ovenfor viser nøkkelverdier for $MSTTR_{W=100}$. Middelverdiene ligger rundt 0,67, og standardavviket i overkant av 0,04. Utvalget er ikke normalfordelt (Shapiro-Wilks normalitetstest: W ≈ 0,956, p < 0,01), hovedsakelig på grunn av en håndfull tekster med lave verdier. De 112 tekstene med verdier over 0,58 er normalfordelt, W ≈ 0,995, p ≈ 0,96, som illustrert i figur 10-32 nedenfor.

[34] Log-TTR og leksikalsk tetthet er begge normalfordelte, men ettersom MSTTR ikke er normalfordelt, bruker jeg Spearmans korrelasjonstest i begge tilfellene for å få sammenlignbare koeffisienter.



Figur 10-32: Tetthetskurve for $MSTTR_{W=100}$. Diagrammet illustrerer avviket fra normalfordeling, samt at variabelen er normalfordelt dersom de åtte laveste verdiene (≤ 0,58) fjernes fra utvalget.

10.4.3 MATTR (Moving Average TTR)

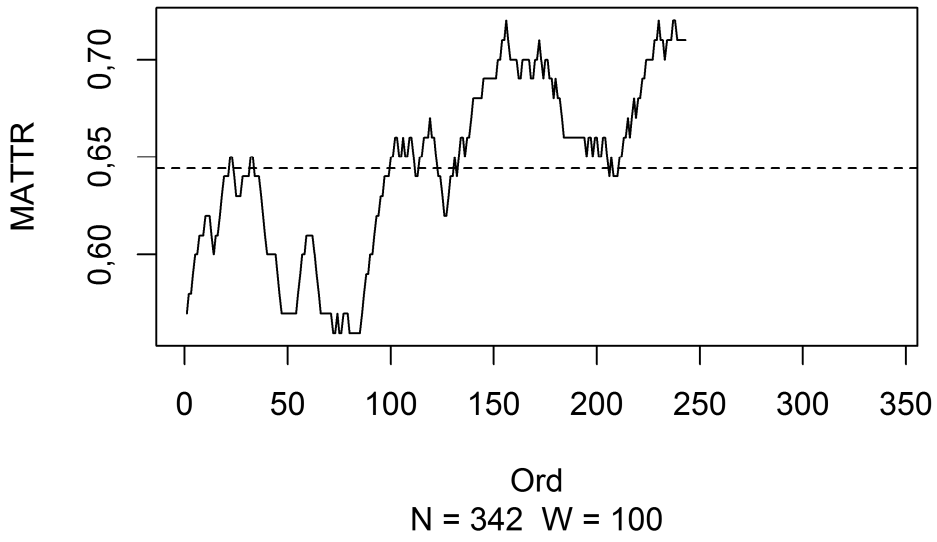
Covington og McFall (2010) hevder å ha hugget over den gordiske knute med en metode de kaller MATTR, *Moving Average TTR*. MATTR er beslektet med MSTTR, men bøter på noen av de uheldige egenskapene ved MSTTR. MATTR er også basert på segmenter eller "vinduer" av teksten, og i dette tilfellet er metaforen vindu kanskje enda mer treffende. MATTR regnes i likhet med MSTTR ut ved hjelp av gjennomsnittlige segmentverdier for TTR, men i MATTR flytter man vinduet bare ett ord utover i teksten for hver gang man beregner TTR. På denne måten blir alle ordene i teksten med i beregningen, og effekten av interaksjoner mellom W og lengden av momentene blir redusert (Covington & McFall, 2010, s. 96): "*MATTR yields a value for every point in the text except for those less than one window length from the beginning*". Denne påstanden stemmer imidlertid ikke overens med slik de selv beskriver prosedyren. Effekten vil være lik ved tekstens begynnelse og slutt, og både begynnelsen og slutten vil være med i beregningen. Imidlertid vil ord i midten av teksten ha W ganger så stor påvirkning på sluttresultatet som både det første og det siste ordet i teksten, dersom teksten er minst $2W$ ord lang. Med $W = 100$ blir altså innvirkningen fra ordene i midten 100 ganger så stor som innvirkningen fra det første og siste ordet. Alle W ordene i hver ende av teksten vil ha mindre påvirkning på MATTR enn alle ordene i midten av teksten. I svært lange tekster, for eksempel på $N \approx 40000$, slik eksempelteksten deres på s. 98 har, vil dette ikke ha noen vesentlig innvirkning på resultatet. (I så lange tekster vil heller ikke det overskytende segmentet for MSTTR ha nevneverdig innvirkning.) For kortere tekster vil dette imidlertid kunne spille en viktig rolle, og det er vanskelig å se at MATTR er noen vesentlig forbedring av MSTTR for relativt korte tekster med hensyn til akkurat denne egenskapen. Når det gjelder visse typer av interaksjon mellom W og avsnittslengder, er imidlertid MATTR åpenbart en forbedring. Ved beregning av MATTR og MSTTR med ulike W for elevtekstkorpuset, er det vesentlig sterkere korrelasjoner mellom MATTR-resultatene enn mellom MSTTR-resultatene. Med verdiene 80, 100 og 120 for W ligger

Pearsons korrelasjonskoeffisienter for MATTR mellom 0,98 og 1,00, mens de for MSTTR ligger mellom 0,92 og 0,94. (Alle oppgitte verdier er avrundede verdier. Utvalgene er ikke normalfordelt, men viser den samme tendensen til skjevfordeling, så en sammenligning av R-verdier er uansett relevant.) Dette viser at tilfeldige variasjoner i TTR-verdiene på bakgrunn av vinduslengde er redusert betraktelig med MATTR i forhold til MSTTR.

Tabell 10-11: Pearsons korrelasjonskoeffisienter mellom MATTR-verdier for ulike verdier av W.

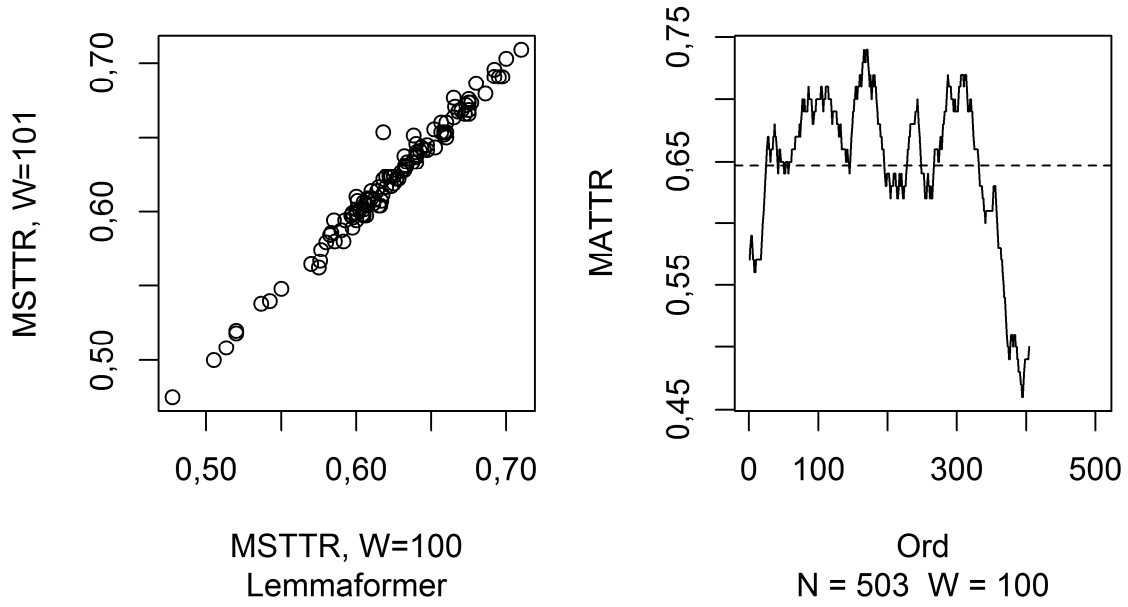
W	**50**	**60**	**80**	**100**	**120**	**140**	**160**	**180**
50	–							
60	0,99	–						
80	0,97	0,99	–					
100	0,93	0,96	0,99	–				
120	0,90	0,93	0,97	0,99	–			
140	0,87	0,91	0,96	0,99	1,00	–		
160	0,86	0,89	0,95	0,98	0,99	1,00	–	
180	0,85	0,88	0,94	0,97	0,98	0,99	1,00	–

Covington og McFall viser dessuten (s. 98) hvordan beregningen av MATTR er nyttig for å vise variasjon i den leksikalske variasjonen gjennom et tekstforløp. Figur 10-33 viser MATTR-verdiene for alle 243 vinduene i en 342 ord lang tekst (A1-235). Diagrammet viser tydelig hvordan variasjonen varierer gjennom tekstforløpet. MATTR-verdien for teksten er middelverdien av MATTR-verdiene for alle vinduene, markert med stiplet linje i figuren.



Figur 10-33: Progressive vindusverdier av MATTR i en tekst (A1-235) av lengde $N = 342$, målt med $W = 100$. Det er særlig markert økning fra x ≈ 80, 130 og 210. Den stiplede linjen markerer gjennomsnittsverdien 0,644 for alle vinduene, altså MATTR for hele teksten.

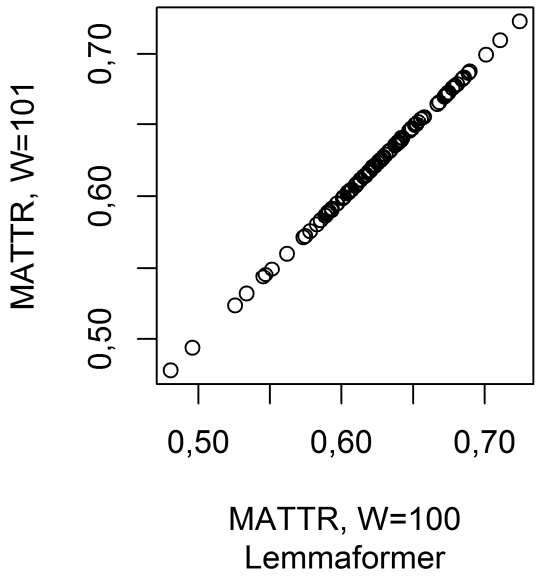
Figur 10-34 nedenfor demonstrerer hvordan MATTR kan belyse individuelle avvik for MSTTR med ulike *W*.[35] I diagrammet til venstre er MSTTR-verdiene for elevtekstkorpuset med $W_1 = 100$ og $W_2 = 101$ plottet mot hverandre. Det er ingen teoretiske grunner knyttet til språklig variasjon til at en økning i segmentlengde på 1 ord skal ha særlig stor innvirkning på resultatet. Likevel ser vi i diagrammet til venstre at særlig én tekst (A1-206) får vesentlig høyere MSTTR med $W_2 = 101$ enn med $W_1 = 100$.



Figur 10-34: Til venstre spredningsdiagram mellom MSTTR for W = 100 og W = 101. Én tekst har særlig stort avvik mellom de to W-verdiene. Til høyre progressiv $MATTR_{W=100}$ for denne teksten. Kurven viser at TTR-verdiene synker mot slutten av teksten. Både MSTTR og MATTR er beregnet på lemmaformer.

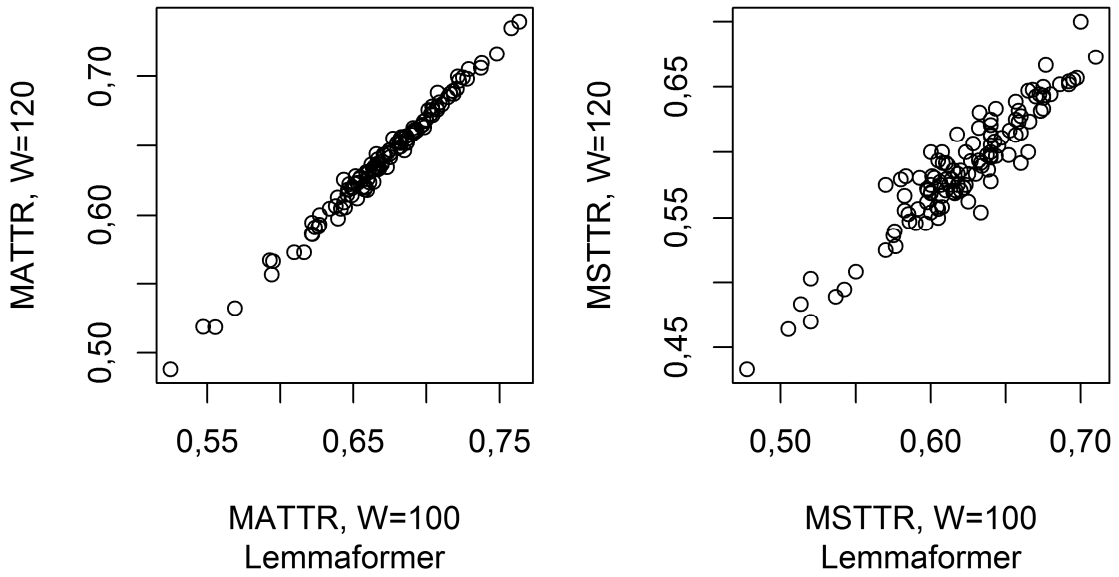
Denne tekstens lengde er N=503, hvilket innebærer at W_1 medfører $k_1 = 5$ segmenter, mens W_2 medfører bare $k_2 = 4$ segmenter, med 99 overskytende ord. Diagrammet til høyre viser progressive vindusverdier for MATTR og demonstrerer at lokal TTR for de siste 100 ordene er vesentlig lavere enn middelverdien for hele teksten, som er $MATTR_{W=100} \approx 0{,}647$, og dette sluttpartiet med markert lavere variasjon enn resten av teksten er det som blir utelatt dersom vi velger $W_2 = 101$. For én tekst er altså konsekvensen temmelig stor, $MSTTR_{W=100} \approx 0{,}618$ og $MSTTR_{W=101} \approx 0{,}653$. For MATTR blir det ingen slike markerte forskjeller for enkelttekster når vi endrer vinduslengden fra $W_1 = 100$ til $W_2 = 101$, og figur 10-35 nedenfor viser også tydelig at korrelasjonen mellom $MATTR_{W=100}$ og $MATTR_{W=101}$ i materialet som helhet er vesentlig større. Sammenligninger mellom ulike *W* med større differanse viser de samme tendensene, men forskjellen mellom MATTR og MSTTR er størst for små differanser i *W*.

[35] I dette eksemplet er lemmaformer brukt som grunnlag for beregningene i stedet for ordformer. Prinsippene som blir illustrert av eksemplet er de samme, men de konkrete tallene vil være andre. Se også diskusjonen i 10.4.5.1.



Figur 10-35: Spredningsdiagram som viser den sterke korrelasjonen mellom MATTR for $W_1 = 100$ og $W_2 = 101$.

MATTR er altså en forbedring av MSTTR på to måter. For det første påvirkes ikke MATTR av tilfeldigheter med hensyn til hvor stor del av teksten som utelates fra beregningen. For det andre reduseres risikoen for visse typer av interaksjon mellom tematiske overganger og segmentgrenser. Figur 10-36 nedenfor viser i hvilken grad MATTR er mindre avhengig av valg av W for to eksempelverdier $W_1 = 100$ og $W_2 = 120$. Tilsvarende forskjeller fremkommer ved andre verdier av W.



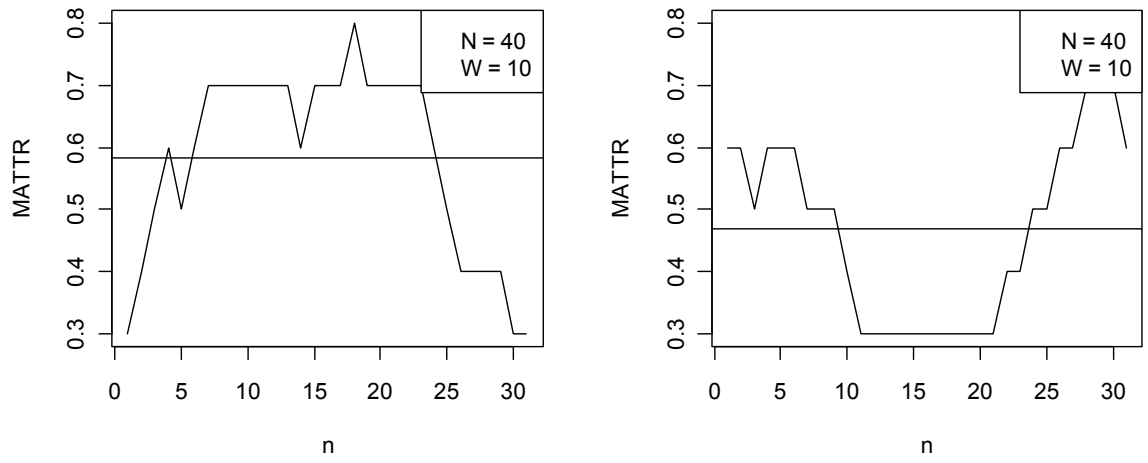
Figur 10-36: Spredningsdiagrammer som viser W-verdienes innvirkning på henholdsvis MATTR (til venstre) og MSTTR (til høyre) for W = 100 og W = 120.

Imidlertid er fortsatt valg av W ikke teoretisk basert, og selv om en endring av verdien fra 100 til 120 har mindre effekt enn for MSTTR, blir forskjellen i W-verdiens effekt på

henholdsvis MATTR og MSTTR mindre når forskjellen mellom *W*-verdier øker. Dette viser at valg av *W* har betydning for hva slags tekstegenskap også MATTR måler.

Dessuten er det uansett slik at MATTR ikke tar tilstrekkelig hensyn til begynnelsen og slutten av teksten. Det er vanskelig å illustrere dette med konkrete tekster, men det kan demonstreres ved simulering.

Jeg har simulert MATTR for stiliserte tekstmodeller av lengde $N = 40$ og med $W = 10$. Tekstmodellene er konstruert slik at de består av 4 like lange deler med potensielt ulike egenskaper. I tekster av type 1 har første og siste del lavere variasjon enn de midterste delene; i tekster av type 2 er det motsatt. Tekstene er konstruert gjennom tilfeldig trekking av symboler fra to alfabet, der det ene alfabetet er av størrelse 3 og det andre av størrelse 10. I tekster av type 1 er innledning og avslutning trukket fra det minste alfabetet og midtdelene fra det største; i tekster av type 2 er det motsatt. TTR og $\text{MSTTR}_{W=10}$ blir den samme for begge typene av tekster, mens $\text{MATTR}_{W=10}$ blir forskjellig.



Figur 10-37: Diagrammer som viser $\text{MATTR}_{W=10}$ for to simulerte tekster av ulik type. Til venstre tekst som har mest variasjon i midten. Til høyre tekst som har mest variasjon i periferien.

Figur 10-37 ovenfor viser to typiske tilfeller av hver teksttype, type 1 til venstre og type 2 til høyre. Vi ser tydelig at variasjonen er lavere i innledning og avslutning i type 1 og lavere i midtdelen i type 2. Den horisontale linjen markerer hele tekstens MATTR, som altså er en gjennomsnittsverdi, og demonstrerer at MATTR er høyere for type 1-teksten til tross for at global TTR-verdi prinsipielt er den samme. Simulering med 1000 tekster av hver type gir middelverdier for $\text{MATTR}_{W=10}$ på henholdsvis $\text{MATTR}_1 \approx 0{,}56$ og $\text{MATTR}_2 \approx 0{,}46$, mens middelverdiene for $\text{MSTTR}_{W=10}$ er $\text{MSTTR}_1 \approx 0{,}4730$ og $\text{MSTTR}_2 \approx 0{,}4735$. Welch' t-test gir $t \approx 68{,}4$ for $\text{MATTR}_{W=10}$ og $t \approx 0{,}31$ for $\text{MSTTR}_{W=10}$. I dette tilfellet gir $W = 30$ selvfølgelig også utslag for MSTTR, $t \approx 40{,}2$, $\text{MSTTR}_1 \approx 0{,}30$, $\text{MSTTR}_2 \approx 0{,}25$. Tallene demonstrerer at både MATTR og MSTTR kan påvirkes av avvikende variasjon i innledning og avslutning for små verdier av *k*.

10.4.4 MOSTTR (Mean Overlapping Segments TTR)

Vi har sett at både MATTR og MSTTR medfører unøyaktigheter knyttet til behandling av tekstavslutning, og at MATTR medfører den samme type unøyaktigheter knyttet til innledning. Vi har dessuten sett at MSTTR påvirkes tilfeldig av små variasjoner i *W* når *N* ligger nær et multiplum av *W*.

Et enkelt grep for å motvirke dette er å øke antall vinduer i MSTTR med 1, slik at alle ordene i teksten blir dekket av (minst) ett vindu. Siden vindusstørrelsen skal være konstant, må det vanligvis innebære en viss overlapp mellom vinduene. Dette medfører at en del ord vil doble sin påvirkning på resultatet. For å motvirke effekten av lokale variasjoner i teksten bør vinduene og dermed overlappingen spres så likt som mulig utover teksten.[36] Da blir det mindre risiko for at en lokal topp eller bunn påvirker resultatet mye. Dessuten vil ingen ord eller segmenter påvirke resultatet mer enn dobbelt så mye som noe annet ord eller segment, i motsetning til MATTR, der mange ord påvirker sluttresultatet opptil *W* ganger så mye som enkelte andre. I de tilfellene der *W* er nøyaktig et multiplum av *N*, eller lik *N*, blir det ingen overlapping, og alle ordene påvirker resultatet like mye. For slike tekster gir MSTTR og MOSTTR samme resultat. For tekster som er bare litt større enn et multiplum av *W*, kan imidlertid overlappingen bli vesentlig, særlig for korte tekster.

Tabell 10-12: Pearsons korrelasjonskoeffisienter mellom MOSTTR-verdier for ulike verdier av W.

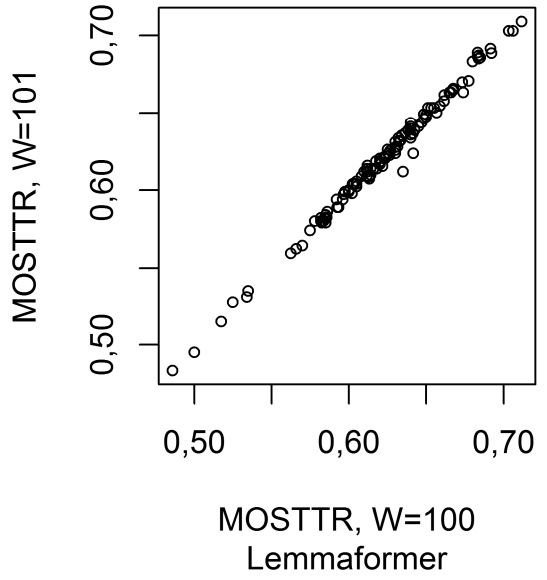
W	**50**	**60**	**80**	**100**	**120**	**140**	**160**	**180**
50	–							
60	0,95	–						
80	0,92	0,94	–					
100	0,94	0,94	0,95	–				
120	0,90	0,92	0,95	0,95	–			
140	0,89	0,90	0,93	0,94	0,96	–		
160	0,88	0,89	0,94	0,94	0,96	0,98	–	
180	0,87	0,88	0,93	0,94	0,95	0,96	0,98	–

Tabell 10-12 ovenfor viser at forskjellene mellom nærliggende *W*-verdier er små, men en god del større enn for MATTR. Imidlertid er korrelasjonen mellom resultater fra *W*-verdier som ligger lengre fra hverandre, mer sammenlignbare. For eksempel er $R \approx 0{,}94$ mellom $MOSTTR_{W=50}$ og $MOSTTR_{W=100}$, mens den for tilsvarende MATTR-verdier er 0,93. Sammenlignet med korrelasjonsverdiene for ordinær MSTTR uten overlapp er også MOSTTR vesentlig bedre, både for små og større differanser av *W*. (Sammenlign med verdiene i figur 10-31.)

[36] Normalt vil ikke alle overlappingene kunne bli like store, ettersom antall ord er diskrete størrelser, men ingen overlapp er mer enn 1 ord lengre enn en annen. Programmet som plasserer vinduene, fordeler overlappingene slik at de største kommer først i teksten. Dette er et tilfeldig valg som ikke vil ha noen vesentlig effekt på resultatet.

MOSTTR med $k + 1$ vinduer er altså noe mer sårbar enn MATTR for små variasjoner i W, men vesentlig mindre sårbar overfor forskjeller i variasjonen mellom tekstens periferi og sentrum, og det er grunnen til at MOSTTR virker som en bedre tilnærming enn MATTR.

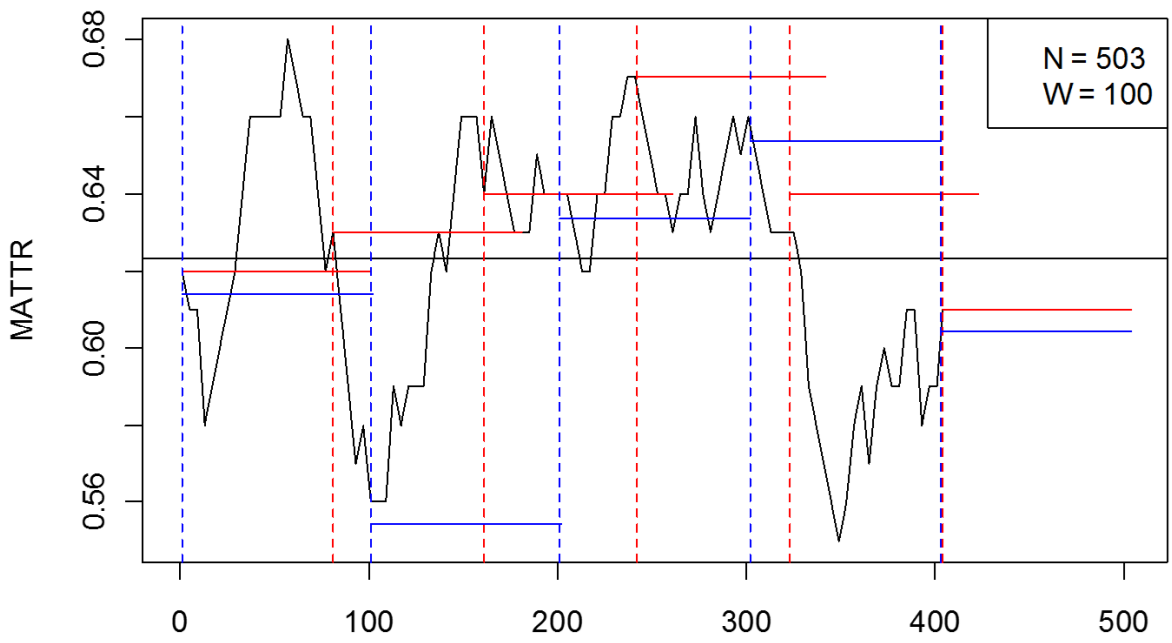
Men i likhet med den opprinnelige MSTTR kan MOSTTR gi vesentlige tilfeldige variasjoner for enkelttekster. For eksempel for A2-245 er $\text{MOSTTR}_{W=100} = 0{,}635$ og $\text{MOSTTR}_{W=101} = 0{,}612$.[37] Denne teksten vises som den med størst avvik i figur 10-38 nedenfor. Dette avviket er likevel vesentlig mindre enn det største avviket mellom $\text{MSTTR}_{W=100}$ og $\text{MSTTR}_{W=101}$.



Figur 10-38: Korrelasjon mellom MOSTTR for $W = 100$ og $W = 101$. To tekster har særlig store avvik. Beregningene er gjort på lemmaformer og ikke ordformer.

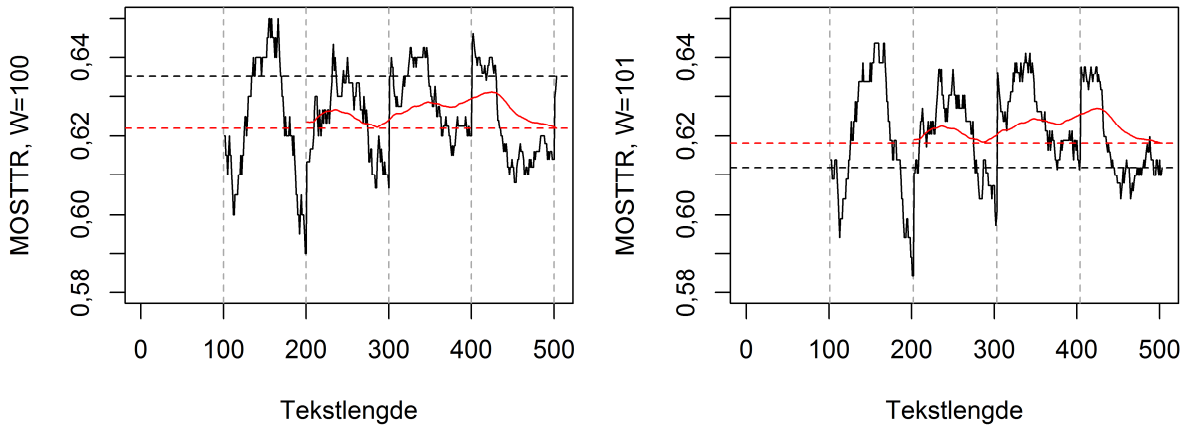
Figur 10-39 nedenfor viser MATTR-verdiene for denne teksten og demonstrerer hvordan et slikt stort avvik kan oppstå for MOSTTR. Siden den ulike vindusstørrelsen medfører ulikt antall vinduer, 6 for $W = 100$ og 5 for $W = 101$, blir målepunktene også plassert på helt ulike posisjoner i teksten. I dette tilfellet har målepunkt nummer 2 spesielt lav verdi for $W = 101$, mens denne bølgedalen i MATTR-diagrammet unngås for $W = 100$. For de andre målepunktene er det ikke spesielt store avvik, men fra figuren kan man se at topp-punktet rundt $x = 60$ og bunnpunktet rundt $x = 350$ er potensielle forstyrrelser i denne teksten som kanskje kunne gi store utslag for andre verdier av W.

[37] I dette eksemplet er lemmaformer brukt som grunnlag for beregningene i stedet for ordformer. Prinsippene som blir illustrert av eksemplet er de samme, men de konkrete tallene vil være andre. Se også diskusjonen i 10.4.5.1.



Figur 10-39: $MATTR_{W=100}$ for teksten A2-245, som har særlig store avvik for MOSTTR mellom $W = 100$ og $W = 101$. Røde streker markerer målepunkter for $MOSTTR_{W=100}$, mens blå streker markerer målepunkter for $MOSTTR_{W=101}$. Beregningene er gjort på lemmaformer og ikke ordformer.

Hvis vi ser på et diagram over progressive MOSTTR-verdier for $W = 100$ og $W = 101$ i figur 10-40 nedenfor, ser vi at vekslingene undervegs i teksten skaper ganske store tilfeldige svingninger i MOSTTR-verdiene. Svingningene er forårsaket av hvilke bølgetopper eller –bunner som blir med to ganger i beregningene.



Figur 10-40: Progressive MOSTTR-verdier for A2-245, som er 503 ord lang. (Beregningene er gjort på lemmaformer og ikke ordformer.) De svarte kurvene markerer progressive MOSTTR-verdier, mens de røde kurvene viser de progressive verdiene for middelverdien av de siste W tekstlengdene.

For $W = 100$ ender beregningen på en bølgetopp, mens den for $W = 101$ ender i en bunn, noe som forklarer det relativt store avviket mellom de to MOSTTR-verdiene for akkurat denne teksten.

Lignende periodiske svingninger kan sees for mange av tekstene, og dette illustrerer at metoden medfører en del tilfeldig variasjon i verdiene. Dette gjelder ikke bare tekster med tekstlengder som er nær et multiplum av W, selv om det er disse tekstene som skiller seg ut i figur 10-38 ovenfor. Denne typen tilfeldig variasjon forstyrrer de systematiske forskjellene som eventuelt finnes i materialet, og svekker styrken i de analysene som skal avdekke disse systematiske mønstrene.

En åpenbar løsning er å utjevne de periodiske forskjellene ved å bruke middelverdien av de W siste verdiene i den progressive MOSTTR. Den røde kurven i diagrammet illustrerer at dette i stor grad fjerner den periodiske variasjonen, og sluttverdien blir temmelig lik for $W = 100$ og $W = 101$. For korpuset som helhet blir det med denne metoden svært små forskjeller mellom to nærliggende W-verdier. Det er imidlertid to ulemper ved denne metoden.

Den første ulempen er knyttet til begynnelsen på den røde kurven. Beregningsmetoden forutsetter at alle tekster er minst $2W$ lange, noe som i elevtekstmaterialet ville nødvendiggjøre en maksimalverdi for W på 91 ord. For kortere tekster kunne riktignok metoden tillempes ved å beregne gjennomsnittsverdien ut fra færre vinduer.

Den andre ulempen er beslektet med argumentasjon jeg brukte for å forkaste MATTR i 10.4.3. Ved å basere MOSTTR på gjennomsnittet av de siste W segmentene, vil de siste W ordene i teksten bli tillagt mindre vekt enn resten, og det aller siste ordet vil bidra med bare $1/W$ så mye som ordene midt i teksten. For teksten i figur 10-40 ser metoden likevel ut til å gi en vesentlig gevinst i forhold til MOSTTR for $W = 100$ eller $W = 101$, men for andre tekster eller andre verdier av W kan ulempene bli store, som vist for MATTR, særlig for korte tekster.

En annen mulighet er å beregne MOSTTR for flere ulike verdier av W og bruke en gjennomsnittsverdi som det endelige TTR-målet, men også denne metoden vil medføre ulik vekting av ulike deler av teksten, spesielt lavere vekting av begynnelse og slutt. Dessuten ville den gjøre spørsmålet om vinduslengde enda mer komplisert.

Konklusjonen av disse mer eksperimentelle forsøkene på å redusere den tilfeldige variasjonen i forskjellige gjennomsnittsberegninger av TTR-mål er å beholde den konseptuelt og matematisk relativt enkle MOSTTR.

10.4.5 MOSTTR-LL (Leksikalsk og lemmaformbasert MOSTTR)

Delkapitlet forklarer og utforsker en forbedret utgave av MOSTTR.

10.4.5.1 Lemmaformer eller ordformer

I tider da datalingvistikken ikke var like godt utviklet, var ordformer den eneste realistiske måten å beregne TTR-baserte mål på i lengre tekster. TTR basert på grafiske ordformer er lette å regne ut og forutsetter *ingen* lingvistisk analyse. Med automatisk lemmatisering og

morfologisk tagging tilgjengelig har man i realiteten et valg mellom å basere analysen av formvariasjon på ordformer, lemmaformer eller leksemer.

Det er to typer av gevinster ved å basere variablene på lemmaformer i stedet for ordformer.

For det første innebærer TTR basert på ordformer i norsk for eksempel at bruk av formene \er\ og \var\ eller \bøker\ og \bøkene\ tilfører teksten like mye målt variasjon som bruk av formene \er\ og \innebar\ eller \bøker\ og \romanene\. Validiteten i målet svekkes ved at det tillegger morfologisk variasjon like mye vekt som leksikalsk variasjon. TTR basert på lemmaformer medfører at man måler tekstens leksikalske variasjon og ikke den morfologiske variasjonen. Man kan argumentere for at også morfologisk variasjon er variasjon, men den er i hvert fall på et lavere nivå og i denne sammenhengen mindre relevant, etter min mening.

For det andre motvirker TTR basert på lemmaformer at homografe ordformer blir regnet som en repetert form. Mange homografe ordformer i norsk er substantiv og presensformer av verb med samme stamme, for eksempel \fester\ eller \fyllekjører\. For disse tilfellene kunne man nok argumentere for at disambiguering av homografien ikke har så stor verdi, ettersom stammen er repetert, akkurat som i bøyningseksemplene over. Men for ordformer som \lyst\, \vekt\ og \bør\ gir det utvilsomt økt validitet å omgå homografien ved å legge de urelaterte lemmaformene \lyst\ og \lys\, \vekt\ og \vekke\, \bør\ og \burde\ til grunn.[38]

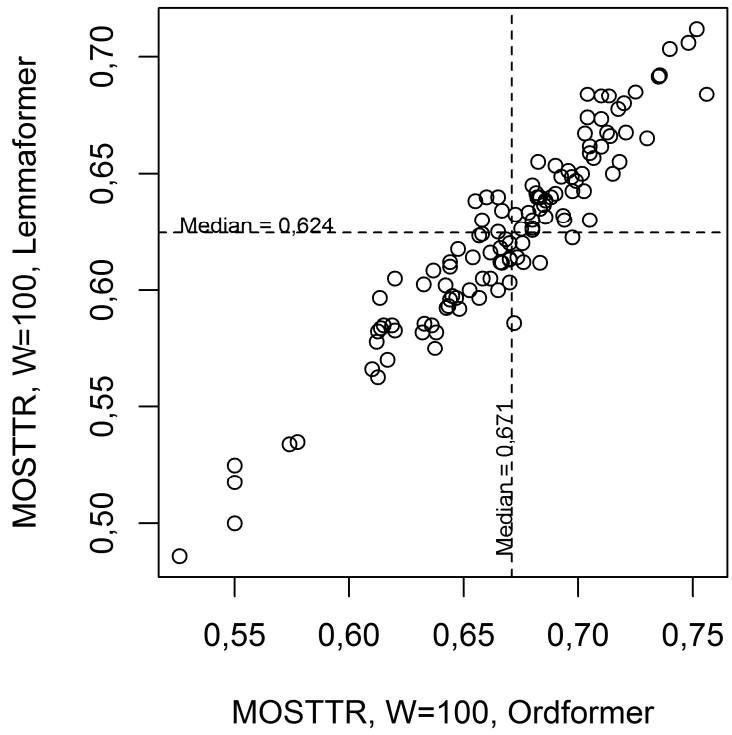
Det er også aktuelt å basere utregningene på leksemer og ikke lemmaformer, særlig der homonyme leksemer tilhører ulike ordklasser, slik det for eksempel kan være for \for\ og \det\. Dette er imidlertid problematisk pga. manglende nøyaktighet i den morfologiske taggingen fra taggerprogrammet. Med det lemmatiseringsprogrammet og de innstillinger som er benyttet til analysene i denne avhandlingen, er det derfor lite aktuelt å forsøke å skille mellom homonyme lemmaformer. Også lemmabaserte analyser er offer for unøyaktigheter i taggingen; for eksempel er 474 forekomster av \bøker\ tilordnet lemmaformen \bøk\. Men omfanget av slike unøyaktigheter er vesentlig mindre enn dersom leksemene skulle være lagt til grunn.

Prinsipielt kunne man tenke seg å gå enda lengre, og benytte rotmorfemer eventuelt sammensetningsledd som enhet, men disse enhetene er ikke direkte tilgjengelige i den automatiske taggingen, og det er derfor ikke aktuelt i denne studien.

Tekstenes TTR-verdier blir selvfølgelig påvirket av hvorvidt utregningene er utført på ordformer eller lemmaformer, men for de norske elevtekstene er ikke forskjellene i verdier så store. Medianverdien for $MOSTTR_{W=100}$ basert på ordformer er 0,67, mens den er 0,62 basert på lemmaformer. Og for en sammenlignende studie av denne typen, har ikke valget særlig store konsekvenser; korrelasjonskoeffisienten mellom de to målene er $R \approx 0{,}95$ (se

[38] Mulige lemmaformer er også \lyse\ og \vek\.

figur 10-41 nedenfor). For andre språk kan konsekvensene av dette metodiske valget være større, eller mindre.



Figur 10-41: Diagram som viser korrelasjonen mellom MOSTTR basert på ordformer og lemmaformer. R ≈ 0,95. Medianen er lavere for lemmaformer enn for ordformer.

Et argument for å holde på en ordformbasert TTR er at dette letter sammenligningen med data fra tidligere undersøkelser. Dessuten blir ikke resultatene påvirket av feilanalyser fra taggerprogrammet – feilanalyser som kan variere fra taggerprogram til taggerprogram. Jeg tror likevel at slike analyser i fremtiden i større grad vil gjøre seg nytte av lingvistiske analyseprogrammer, og at et lemmabasert mål er et prinsipielt og validitetsmessig riktigere mål. Når de grammatiske analyseprogrammene blir bedre, bør man også gå et skritt videre og benytte leksemer. I denne avhandlingen har jeg valgt å basere analysene på lemmaformer.

10.4.5.2 Leksikalsk TTR

Alle typer TTR-baserte variasjonsmål som vi har snakket om så langt, påvirkes av leksikalsk tetthet. Lav leksikalsk tetthet påvirker TTR-baserte variasjonsmål negativt rett og slett fordi klassen av grammatiske ord er svært mye mindre enn klassen av leksikalske ord, og det dermed er mindre variasjon blant grammatiske ord enn blant leksikalske ord. MOSTTR-$L_{W=100}$ korrelerer betydelig med leksikalsk tetthet, rho ≈ 0,35. (MOSTTR-$L_{W=100}$ er ikke normalfordelt, hovedsakelig på grunn av 6 tekster med lave verdier, og jeg har derfor benyttet Spearmans korrelasjonstest.)

Fordi det er ønskelig med variabler som isolerer mest mulig atomiske egenskaper ved tekstene, bør leksikalsk tetthet beregnes som en variabel for seg, og dens innvirkning på

målet for leksikalsk variasjon bør reduseres eller fjernes. Dette oppnås ved å beregne TTR bare av leksikalske ord. Som jeg påpeker og drøfter i kapittel 9 ovenfor, er begrepene leksikalske ord og grammatiske ord prototypiske begreper uten entydige avgrensninger, og det er derfor prinsipielt umulig å trekke en klar og entydig skillelinje mellom disse to kategoriene. I denne avhandlingen definerer jeg leksikalske ord slik det er forklart i kapittel 9.

Siden den leksikalske tettheten i korpuset har middelverdier rundt 0,4, blir tekstene i korpuset jevnt over under halvparten så lange dersom alle funksjonsordene fjernes. Dermed er heller ikke de samme *W*-verdiene aktuelle. Jeg har valgt $W = 50$ for MOSTTR-LL, leksikalsk lemmabasert MOSTTR; da er *W* fortsatt vesentlig kortere enn den korteste teksten, som har 84 leksikalske ord. Med dette grepet er korrelasjonen mellom TTR-målet og leksikalsk tetthet så godt som eliminert, rho ≈ 0,042, og jeg har dermed oppnådd to uavhengige leksikalske variabler som representerer to ulike egenskaper ved tekstene.

Korrelasjonen mellom MOSTTR basert på alle lemmaformer og MOSTTR basert på bare leksikalske lemmaformer er sterk, rho ≈ 0,74, men ikke ekstremt sterk. Det er åpenbart en forskjell i hva de to variablene måler.

10.4.5.3 Analyse

På bakgrunn av diskusjonen så langt i dette delkapitlet har jeg valgt å bruke MOSTTR-varianten av det lokale TTR-mål. Analysen er basert på lemmaformer og kun de leksikalske ordene i tekstene, og segmentlengden er valgt til $W = 50$. Nøkkelverdiene for $\text{MOSTTR-LL}_{W=50}$ er presentert i tabell 10-13 nedenfor.

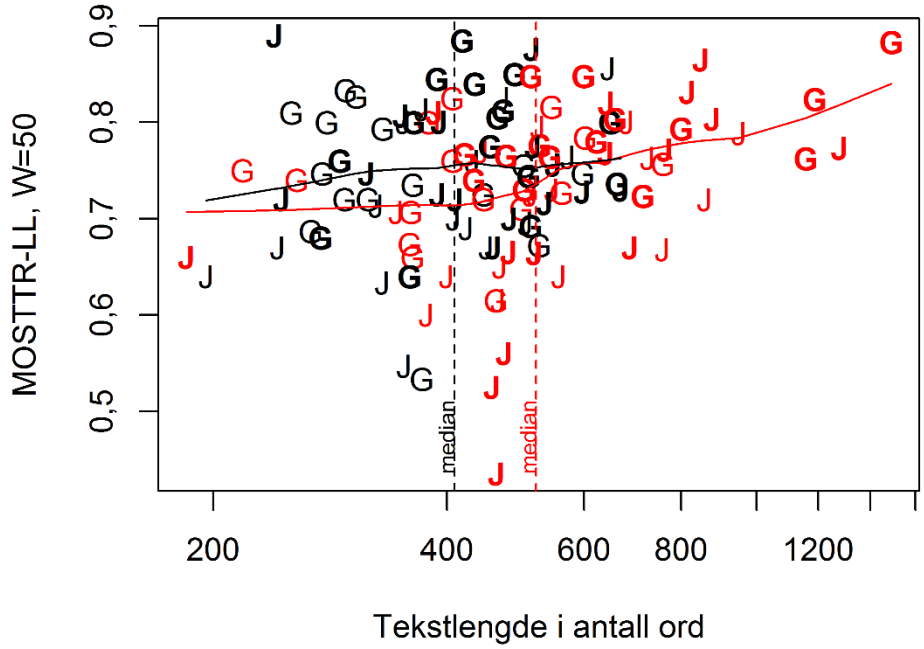
Tabell 10-13: Nøkkelverdier for MOSTTR-LL, $W = 50$

	middelverdi	median	sd	min	maks
Total	**0,742**	**0,748**	**0,080**	**0,435**	**0,890**
Hånd	0,748	0,747	0,075	0,533	0,890
Tast	0,735	0,758	0,085	0,435	0,883
Middels	0,726	0,726	0,072	0,533	0,856
Sterk	0,758	0,767	0,085	0,435	0,890
Gutt	0,760	0,761	0,066	0,533	0,885
Jente	0,723	0,725	0,089	0,435	0,890

Middelverdiene ligger i underkant av 0,75, med standardavvik rundt 0,08. Utvalget er ikke normalfordelt, $W \approx 0{,}961$, $p < 0{,}01$ med Shapiro-Wilks normalitetstest, hovedsakelig på grunn av 5 tekster med lave verdier, slik det går fram av figur 10-42. Alle disse 5 tekstene er "Bøker eller data"-tekster. Uten disse 5 tekstene er utvalget normalfordelt, $W \approx 0{,}988$, $p \approx 0{,}40$.

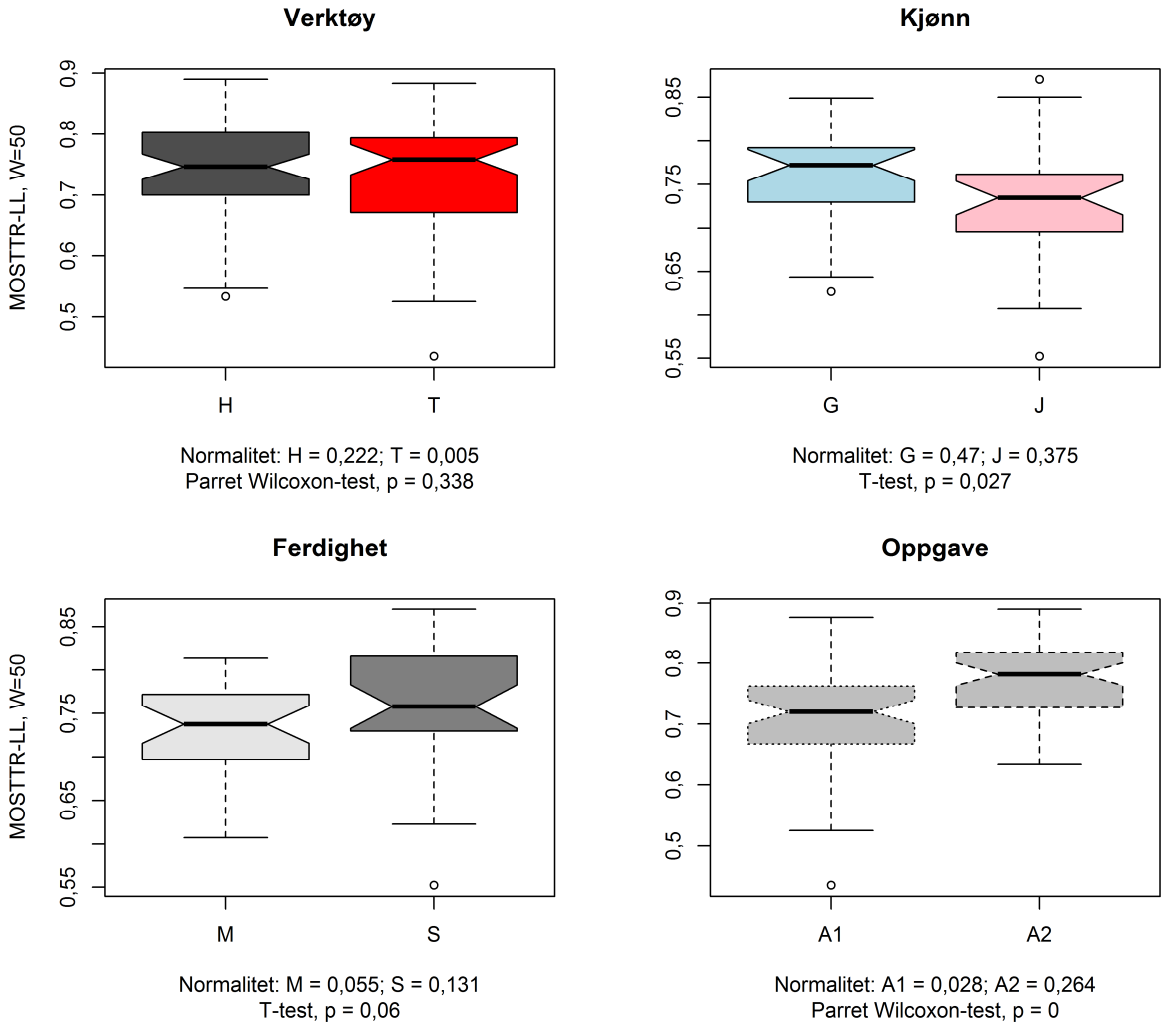
Figur 10-42 nedenfor viser også at det er en svak korrelasjon med tekstlengde, rho ≈ 0,22, $p < 0{,}05$. Denne korrelasjonen har ingenting med den matematiske sammenhengen mellom TTR og tekstlengde å gjøre, men er snarere et tegn på at elever som skriver langt, også skriver med noe større leksikalsk variasjon. Sammenhengen mellom tekstlengde og

leksikalsk variasjon kan være uttrykk for flere faktorer, blant annet elevens generelle skriveferdighet, tekstens momentrikdom og elevens motivasjon. Men det kan også være uttrykk for svakere koherens i de lengre tekstene.



Figur 10-42: Spredningsdiagram som viser sammenhengen mellom MOSTTR-LL$_{W=50}$ og tekstlengde, rho ≈ 0,22, p ≈ 0,017.

Figur 10-43 nedenfor viser at verktøy ikke har noen innvirkning på variabelen, mens guttene også med dette variasjonsmålet har en del høyere leksikalsk variasjon enn jentene. De sterke elevene skriver med noe høyere variasjon enn de middels sterke elevene. Variasjonen er dessuten klart høyest i "Ungdomsfylla"-tekstene. Korrelasjonen mellom håndtekster og tastetekster er moderat, rho ≈ 0,33, N = 2x60.



Figur 10-43: MOSTTR-LL$_{W=50}$ og skriveverktøy, kjønn, ferdighet og oppgave

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, antall interaksjonsnivåer begrenset til 2:

(121) `lm(lexD$MOSTTR_LL.50~(kjønn+ferdighet+lengde+forskjell)^2)`

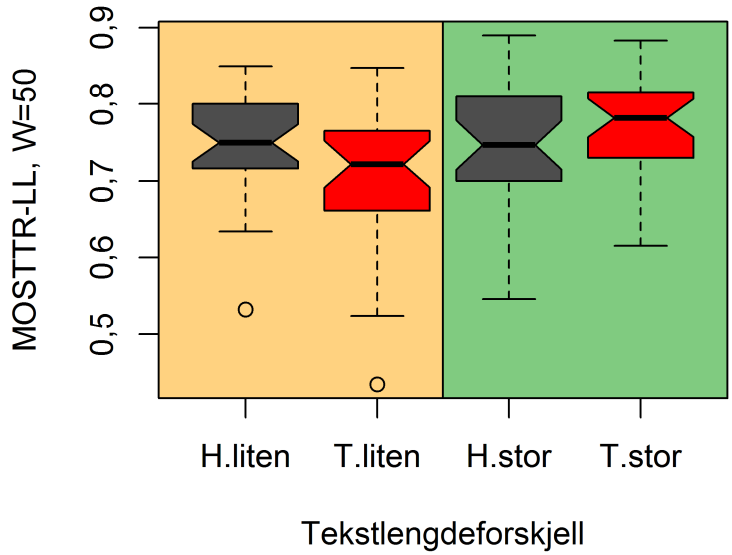
Den maksimale modellen over ble redusert til følgende minimale adekvate modell:

(122) `lm(formula = lexD$MOSTTR_LL.50 ~ forskjell)`

```
             Df Sum Sq Mean Sq F value Pr(>F)
forskjell     1 0.0666 0.06659   9.267 0.0035 **
Residuals    58 0.4168 0.00719
---
Multiple R-squared:  0.1378,    Adjusted R-squared:  0.1229
F-statistic: 9.267 on 1 and 58 DF,  p-value: 0.003504
```

Tekstlengdeforskjell er den eneste signifikante prediktoren, $F \approx 9{,}3$, $p < 0{,}01$, $d \approx 0{,}80$, noe som gjør at denne variabelen har blant de sterkeste effektene i hele undersøkelsen. `Gvlma` (se 7.2.2.4) viser at premissene for anova-analysen er oppfylt. (Se appendiks A4.)

Figur 10-44 viser at forskjellen består i at elever med liten forskjell i tekstlengde har mindre variasjon i tastetekstene, mens det er motsatt for de elevene som skriver mye lengre på tastatur. Det er ingen forskjell mellom utvalgene i håndtekstene.



Figur 10-44: Resultatet av anova-analysen på MOSTTR-$LL_{W=50}$. Elever med liten tekstlengdeforskjell har lavere variasjon i tastetekstene. Elever med stor tekstlengdeforskjell har noe høyere variasjon i tastetekstene.

10.4.5.4 Oppsummering

Segmentbaserte TTR-mål representerer først og fremst en lokal, språklig variasjon og fungerer dermed som et supplement til de logaritmetransformerte, globale TTR-målene. Korrelasjonen er sterk mellom de to typene, representert ved MOSTTR-$LL_{W=50}$ og log-$TTR_{1,3}$, $R \approx 0{,}85$, $N = 2x60$, men ikke sterkere enn at vi kan regne med at bidrar med to ulike dimensjoner av leksikalsk variasjon.

Utgangspunktet for de segmentbaserte målene var FSTTR, som i større grad måler en global variasjonsegenskap ved tekstene, riktig nok begrenset av forkastingen av deler av hver tekst. I jakten på en variasjonsvariabel som tar hensyn til hele teksten, endte vi opp med en variabel som gjør nettopp det, men som gjennom å segmentere teksten i kortere segmenter i større grad måler den lokale variasjonen. Det faktum at både FSTTR og de variablene som bygger på gjennomsnittsverdier av kortere segmenter, korrelerer med tekstens lengde, er med å underbygge argumentasjonen om at også de globale variasjonsvariablene bør korrelere med tekstlengde, og at de dermed må justeres tilstrekkelig til at de får en positiv korrelasjon.

En vesentlig komponent i verdiene av TTR-baserte variabler med utgangspunkt i ord- eller lemmaformer skriver seg fra tekstens leksikalske tetthet. Ved å fjerne funksjonsordene fra

tekstene og beregne variasjonen utelukkende blant de leksikalske ordene oppnår man et leksikalsk variasjonsmål som er uavhengig av leksikalsk tetthet. Det nesten totale fraværet av korrelasjon mellom leksikalsk tetthet og MOSTTR-LL$_{W=50}$ viser at vi her har fått to nærmest uavhengige dimensjoner av leksikalske tekstegenskaper. Også for de transformasjonsbaserte TTR-variablene kunne man fjerne funksjonsordene og oppnå en rent leksikalsk variabel, men dette har jeg ikke gjort.

Ved å basere utregningene på lemmaformer heller enn ordformer har jeg dessuten fjernet den morfologiske variasjonen fra variabelen. Også de transformasjonsbaserte TTR-variablene kunne gjerne vært basert på lemmaformer, men jeg beholdt ordformen som grunnleggende enhet for disse variablene. Ved å beholde ett variasjonsmål basert på ordformer og ett variasjonsmål basert på lemmaformer, står vi igjen med i hvert fall én variabel der også den morfologiske variasjonen er representert.

Til slutt vil jeg nevne korrelasjonen mellom MOSTTR-LL$_{W=50}$ og gjennomsnittlig ordlengde, og den enda sterkere korrelasjonen med gjennomsnittlig ordlengde i de leksikalske ordene ($R \approx 0{,}53$). Dette viser at disse enkle målene fanger vesentlige leksikalske egenskaper, trolig også egenskaper som ikke fanges av noen av de andre leksikalske variablene vi så langt har diskutert.

10.5 Andre variasjonsmål

Så langt i kapitlet har jeg drøftet ulike variasjonsmål som bygger på TTR. Til slutt skal jeg se ganske kort på to andre typer mål på leksikalsk variasjon, nemlig statistikk knyttet til unike ord (hapax legomena) og en entropibasert utregningsmåte.

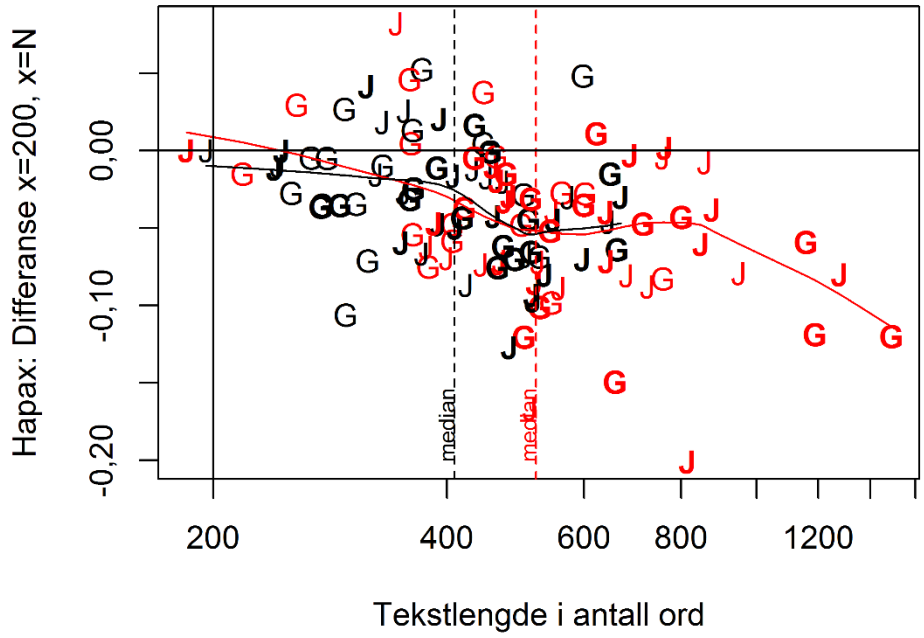
10.5.1 Hapax legomena

Hapax legomena er en betegnelse på de ordene som er unike i en tekst. Kvantitative egenskaper knyttet til disse ordene kan åpenbart gjenspeile relevante egenskaper ved teksten; vi så for eksempel i figur 10-3 på side 172 at kriminalromanen hadde vesentlig flere unike ord enn begge de to andre tekstene.

(Hultman & Westman, 1977, s. 60) hevder at antall hapax legomena per antall ordtyper er relativt konstant i tekster av lengde mellom 200 og 2000 ord, som nesten alle tekstene i elevtekstkorpuset er. Figur 10-47 viser imidlertid at det er en svak negativ korrelasjon mellom dette frekvensmålet for unike ord og tekstlengde.

At det er en korrelasjon mellom hapax-tettheten og tekstlengde, trenger imidlertid ikke bety at hapax-tettheten ikke er konstant for hver forfatter – eller for hver forfatter innenfor en skrivesituasjon. Korrelasjonen med lengden trenger ikke bety annet enn at elever som skriver lengre også har en annen hapax-tetthet. Imidlertid er det påfallende at dette målet for variasjon viser motsatt tekstlengdetendens av de lengdeuavhengige TTR-målene, og jeg synes det er grunn til å tvile på om Hultman og Westman har rett i at variabelen er konstant for disse tekstlengdene.

Hvis man beregner differansen mellom hapax-tetthet for alle tekstene beskåret til n = 200 og hapax-tettheten for tekstene i sin fulle lengde, ser man at hapax-tettheten er høyere ved n = 200 for de aller fleste tekstene, og at tendensen forsterkes med tekstlengde (figur 10-45). Regresjonskurvene er dessuten påfallende sammenfallende for håndtekster og tastetekster.



Figur 10-45: Differanse mellom hapaxtetthet ved n = 200 og ved tekstens slutt. Diagrammet viser en sterk overvekt av negative verdier, samt at tendensen er synkende for økende tekstlengde.

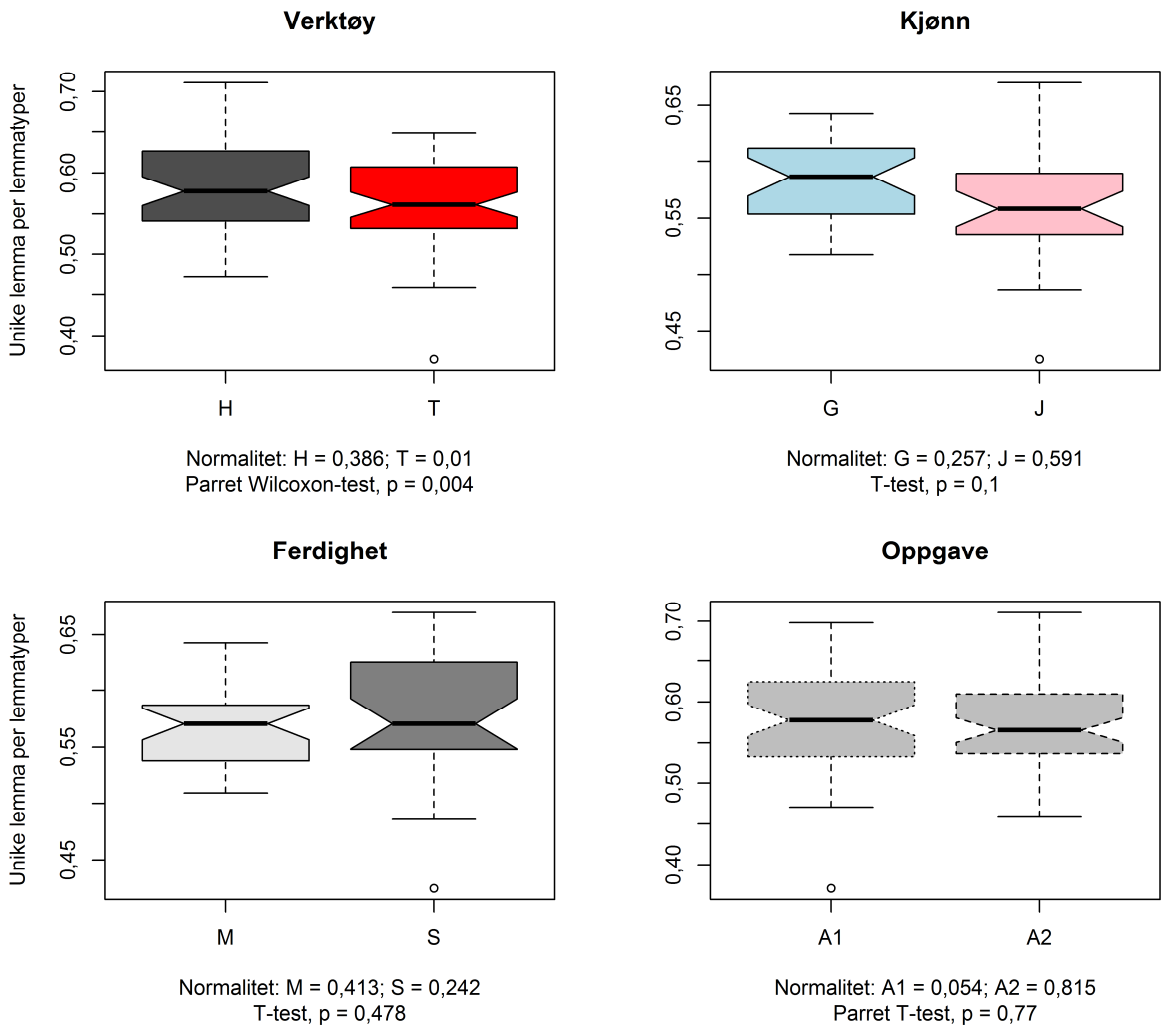
Dette tyder på at sammenhengen mellom hapax-tettheten og tekstlengde er av matematisk karakter og ikke henger sammen med disse elevenes skrivemønstre. Dette gjør variabelen mindre interessant i denne analysen, og jeg behandler den derfor bare ganske knapt. Jeg følger konklusjonene fra diskusjonen i 10.4.5.1 og benytter lemmaformer som enhet for analysen.

10.5.1.1 Deskriptiv analyse

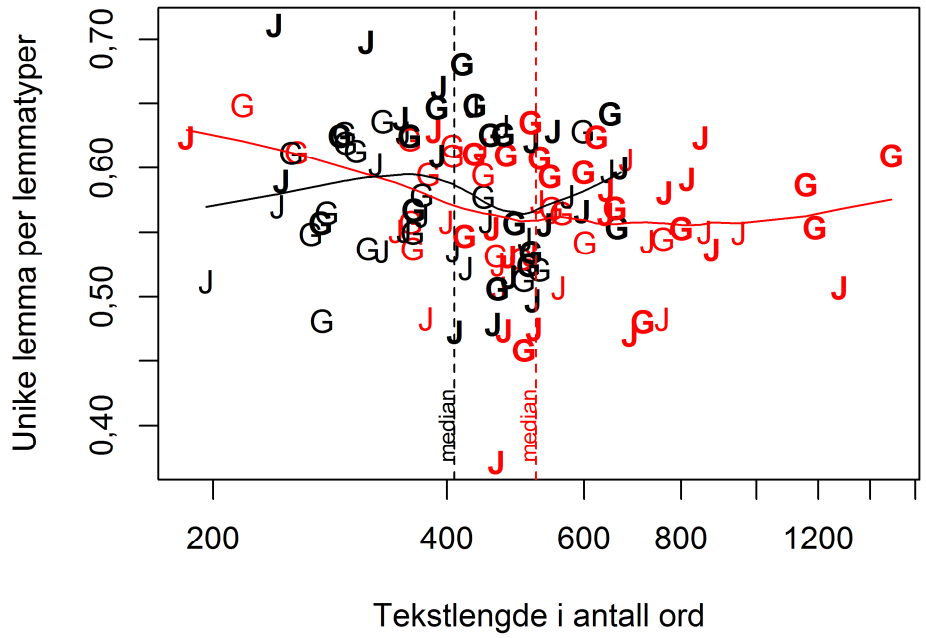
Tabell 10-14 viser at gjennomsnittsverdiene ligger litt under 0,6, med 0,05 som standardavvik.

Tabell 10-14: Nøkkeltall for hapax legomena per antall ordtyper, lemmaformbasert.

	middelverdi	median	sd	min	maks
Total	**0,571**	**0,567**	**0,055**	**0,371**	**0,710**
Hånd	0,583	0,578	0,056	0,473	0,711
Tast	0,560	0,561	0,053	0,371	0,649
Middels	0,567	0,561	0,043	0,481	0,649
Sterk	0,575	0,585	0,066	0,371	0,711
Gutt	0,581	0,583	0,048	0,459	0,681
Jente	0,561	0,558	0,061	0,371	0,711



Figur 10-46: Unike lemmaformer per lemmatyper fordelt etter fire faktorer.



Figur 10-47: Tetthet av hapax legomena mot tekstlengde. Rho ≈ –0,23.

Figur 10-47 viser en tendens til negativ korrelasjon med tekstlengde, rho ≈ –0,23. Utvalget er normalfordelt ifølge Shapiro-Wilks normalitetstest (W ≈ 0,982, p ≈ 0,11), men har såpass store skjevheter at jeg har valgt å holde meg til Spearmans korrelasjonstest. Diagrammet viser blant annet ei sterk jente med en ekstremt lav verdi i tasteteksten sin.

10.5.1.2 Variansanalyse

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, antall interaksjonsnivåer begrenset til 2:

(123)
```
lm(lexD$hapaxF ~ (kjønn+ferdighet+lengde+forskjell)^2)
```

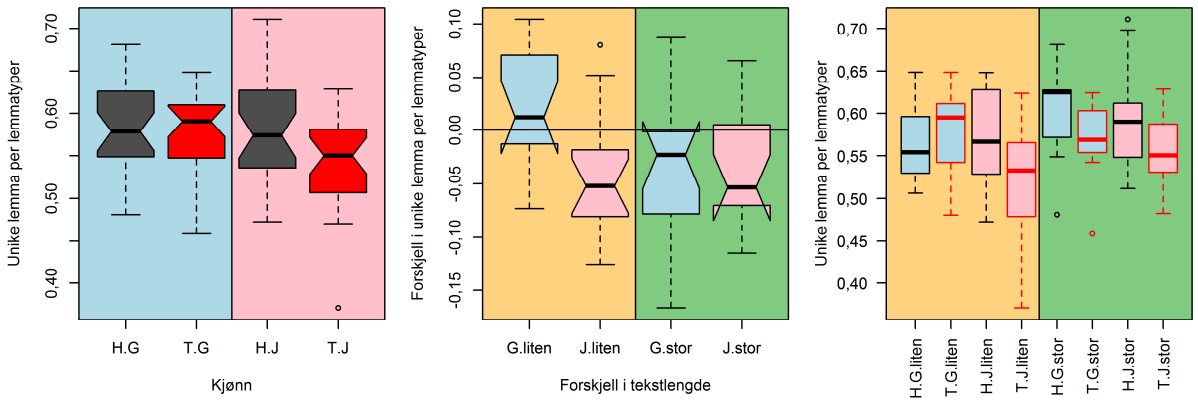
Reduksjon av den maksimale modellen resulterer i følgende minimale adekvate modell:

(124)
```
lm(formula = lexD$hapax.lemmaF ~ kjønn + forskjell +
kjønn:forskjell)

                 Df  Sum Sq  Mean Sq F value Pr(>F)
kjønn             1 0.01489 0.014891   4.743 0.0336 *
forskjell         1 0.00707 0.007074   2.253 0.1389
kjønn:forskjell   1 0.01550 0.015504   4.939 0.0303 *
Residuals        56 0.17580 0.003139
---
Multiple R-squared:  0.1757,    Adjusted R-squared:  0.1315
F-statistic: 3.978 on 3 and 56 DF,  p-value: 0.0122
```

Modellen er svakt signifikant, F ≈ 3,98, p < 0,05. Kjønn (d ≈ 0,55) og interaksjonen mellom kjønn og forskjell er svakt signifikante prediktorer, p < 0,05. Tukeys HSD-test gir som resultat at forskjellen mellom gutter med liten tekstlengdeforskjell og jenter med liten tekstlengdeforskjell er svakt signifikant, p < 0,05, d ≈ 1,23. (Se appendiks A5. To andre interaksjoner er nær signifikante.) `Gvlma` (se 7.2.2.4) viser at premissene for variansanalysen er oppfylt. (Se appendiks A4.)

Diagrammet til venstre i figur 10-48 viser at jentene har lavere hapax-tetthet i tastetekstene, mens det ikke synes å være noen forskjell i guttenes tekster. Diagrammet i midten avslører imidlertid at bildet er mer nyansert. Gutter med liten forskjell i tekstlengde har høyere hapax-tetthet i tastetekstene, mens tendensen hos de andre guttene og hele jentegruppen er lavere hapax-tetthet i tastetekstene. Diagrammet til høyre viser at forskjellen mellom de to segmentene av gutter kanskje først og fremst er at verdiene blant guttene med stor tekstlengdeforskjell er høyere i håndtekstene, mens verdiene i de to segmentene er ganske like i tastetekstene. Det er kanskje også en tendens til høyere verdier generelt blant guttene med stor tekstlengdeforskjell enn blant jentene.



Figur 10-48: Unike lemmaformer per lemmatyper, resultater fra anova-analysen. Til venstre interaksjon mellom skriveverktøy og kjønn. I midten effekten av interaksjonen av kjønn og tekstlengdeforskjell på differansen av verdiene. Til høyre interaksjonen mellom kjønn, tekstlengdeforskjell og skriveverktøy.

Bildet er altså komplisert. Den generelle tendensen til lavere verdier i tastetekstene som går fram av figur 10-46 skyldes jentene og guttene med stor forskjell i tekstlengde. Den signifikante kjønnseffekten som går fram av anova-analysen og det venstre diagrammet i figur 10-48, skyldes først og fremst det segmentet av guttene som har liten tekstlengdeforskjell. Forskjellen mellom gutter med liten tekstlengdeforskjell og jenter med liten tekstlengdeforskjell skyldes først og fremst at disse guttene har høyere verdier i tastetekstene enn de tilsvarende jentene.

10.5.1.3 Diskusjon

Kjønnseffekten for tetthet av hapax legomena ligner på den kjønnseffekten vi har sett for en del andre leksikalske variabler, som gjennomsnittlig ordlengde, gjennomsnittlig lengde av leksikalske ord og *log-TTR*$_{1,3}$, nemlig at jenter har lavere verdier i tastetekstene, mens

guttene som gruppe er lite påvirket av skriveverktøy. Ingen av de andre leksikalske variablene har effekter som innebærer interaksjoner av kjønn og tekstlengdeforskjell, men flere har forskjellige typer effekter av tekstlengdeforskjell. Leksikalsk tetthet er høyere i tastetekstene for sterke elever som har liten tekstlengdeforskjell, altså en effekt som er en parallell til effekten på hapax-variabelen, dersom vi antar at høyere verdier er et uttrykk for mer planlagt, redigert tekst. For de TTR-baserte målene *log-TTR*$_{1,3}$ og MOSTTR-LL$_{W=50}$ er effekten den motsatte, nemlig at elevene med liten tekstlengdeforskjell har lavere verdier i tastetekstene, mens elevene med stor tekstlengdeforskjell har høyere verdier i tastetekstene. Igjen er det et premiss for tolkningen at høyere leksikalsk variasjon faktisk er et uttrykk for mer redigert tekst; denne tolkningen av de TTR-baserte målene er så langt usikker.

Hapax legomena-tetthet korrelerer positivt (N = 2x60) med alle de nevnte leksikalske variablene, ordlengde ($R \approx 0{,}42$), leksikalsk ordlengde ($R \approx 0{,}41$), *log-TTR*$_{1,3}$ ($R \approx 0{,}66$) og MOSTTR-LL$_{W=50}$ ($R \approx 0{,}58$), og dessuten med leksikalsk tetthet ($R \approx 0{,}28$).[39] Dette tyder på at de i en viss forstand er uttrykk for beslektede tekstlige egenskaper, selv om de ikke nødvendigvis er det på en dimensjon som har med spontanitet kontra redigerthet å gjøre.

Jeg peker i 10.6 på det mangelfulle i TTR-baserte variabler som mål på leksikalsk variasjon, og kritikken er blant annet knyttet til at TTR ikke tar hensyn til hvordan ordeksemplarene er fordelt på ordtypene. Å undersøke unike lemmaformer i tekstene er et forsøk på å svare på denne kritikken ved at dette er en variabel som påvirkes positivt av den type variasjon som fremkommer gjennom oppfinnsomme synonymer, men riktignok også kanskje også av mangel på utdyping av tema. Jeg tror antall unike lemmaformer i en tekst har potensial som del av en leksikalsk variabel. Jeg er imidlertid mer i tvil om hvorvidt målestokken som Hultman og Westman (1977, s. 60) kom fram til, er en validitetsmessig fornuftig målestokk, og jeg synes det er vanskelig å hevde at andelen lemmaformtyper som er unike i teksten, representerer en intuitivt lettfattelig egenskap ved teksten. Variabelen er dessuten sårbar for interaksjon med antall ordtyper i teksten.

At variabelen dessuten korrelerer tydelig negativt med tekstlengde, både målt i korpuset som helhet og i den enkelte tekst, understreker variabelens validitetsproblemer, og det gjør at resultatene ovenfor må tolkes med forsiktighet, særlig effekten av tekstlengdeforskjell, men også kjønnsforskjellen. Å videreutvikle variabelens korrelasjon med tekstlengde ved hjelp av en segmentbasert løsning som for TTR er ikke en farbar veg, ettersom vi neppe er særlig interessert i antall lemmaformer som er unike innenfor kortere segmenter. Å bruke studier av progressive verdier til å justere variabelen etter tekstlengde er imidlertid mulig å forestille seg, men variabelen ville i så fall også miste enda mer av sin intuitive fortolkning.

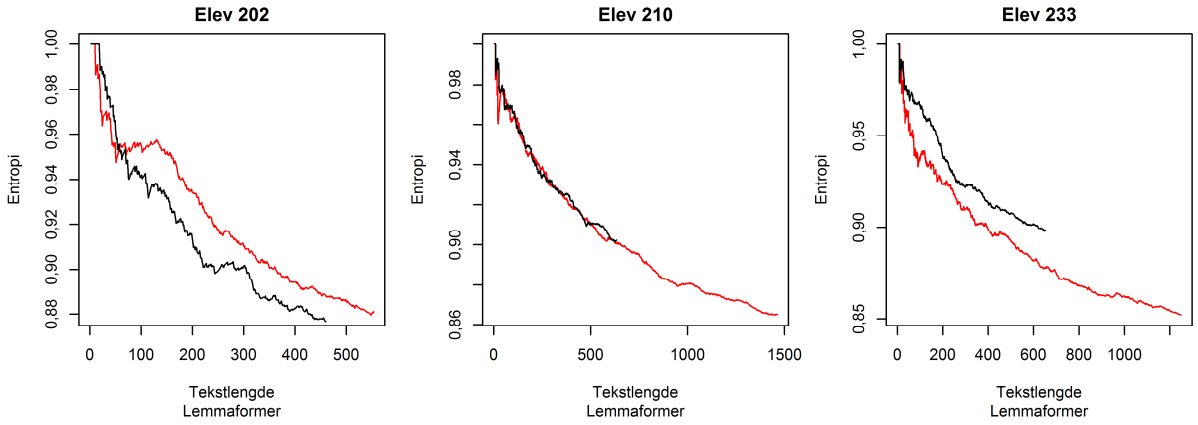
[39] Ikke alle disse variablene er normalfordelt, men jeg bruker Pearsons korrelasjonstest for å få sammenlignbare korrelasjonskoeffisienter.

Konklusjonen er at jeg ikke bruker fordeling av unike lemmaformer i videre analyser i denne studien.

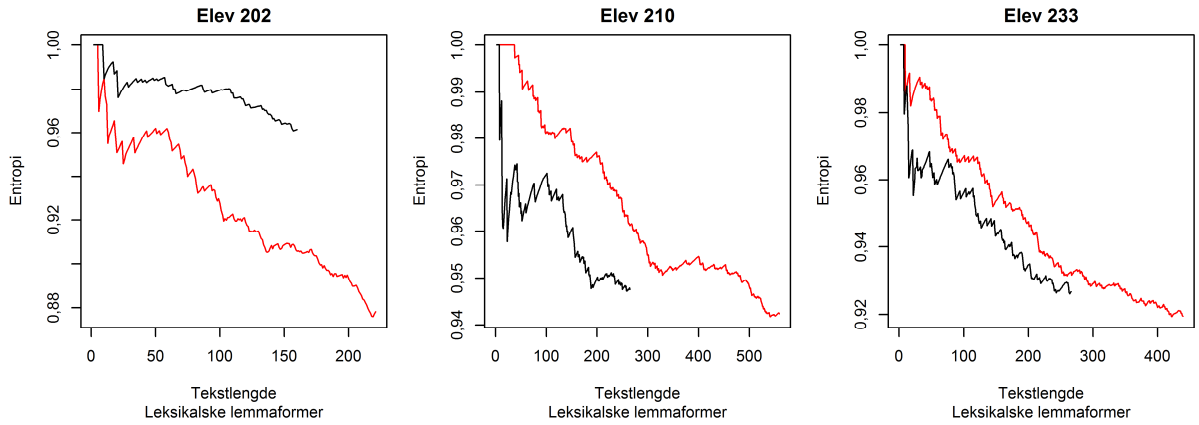
10.5.2 Entropi

Entropi (H) av ordtypers frekvens kunne potensielt være en variabel som fanger mer av *variasjonen* i ordtilfanget (se 2.1.2) enn bare forholdstall mellom ordtyper og ordeksemplarer.

Diagrammene i figur 10-49 nedenfor over progressive entropi-verdier for tekstene til tre elever viser imidlertid at entropi av ordtypefrekvenser basert på lemmaformer *ikke* er tekstlengdeuavhengig, men entydig sterkt fallende gjennom tekstlengden. Også for frekvenser av leksikalske lemmaformtyper (figur 10-50) er tendensen udiskutabel, men synes ikke like sterk. I elevtekstkorpuset er det svært sterk korrelasjon mellom (logaritmetransformert) tekstlengde og entropi basert på lemmaformtyper ($R \approx -0{,}84$, $N = 2x60$), mens tilsvarende tendens for bare de leksikalske lemmaformtypene er svak ($R \approx -0{,}22$, $N = 2x60$). Entropi for lemmaformtyper er normalfordelt ifølge Shapiro-Wilks normalitetstest ($W \approx 0{,}993$, $p \approx 0{,}73$, $N = 2x60$), mens det for leksikalske lemmaformtyper er en liten håndfull lave verdier som forstyrrer normaliteten ($W \approx 0{,}961$, $p \approx 0{,}002$, $N = 2x60$), og korrelasjonskoeffisienten må tolkes med varsomhet. At korrelasjonen er mye svakere for denne variabelen, er imidlertid hevet over tvil.



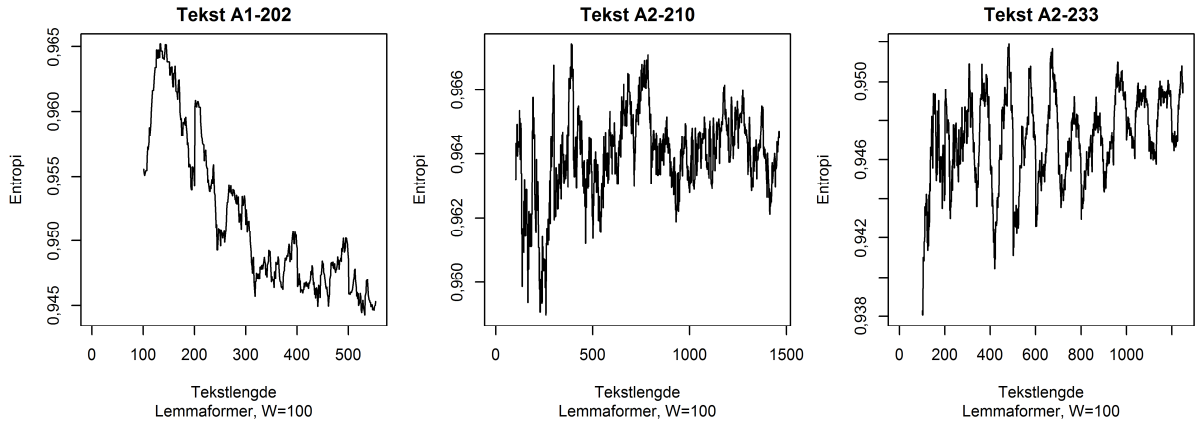
Figur 10-49: Entropi av ordtypers frekvens i tekster av tre ulike elever, for hånd (svart) og med tastatur (rødt), basert på lemmaformer. Hver kurve viser progressive verdier av entropi etter hvert som teksten "vokser" seg lengre.



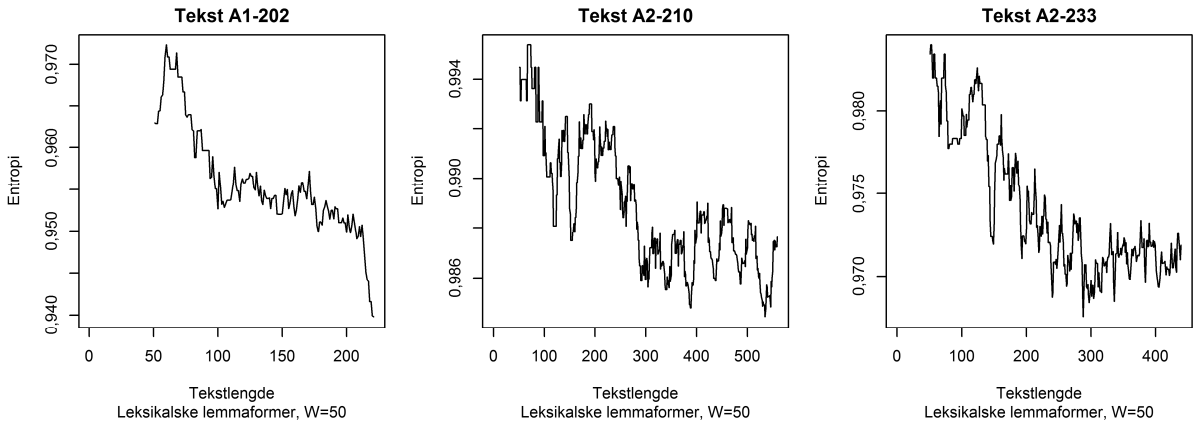
Figur 10-50: Entropi av ordtypers frekvens i tekster av tre ulike elever, for hånd (svart) og med tastatur (rødt), basert på lemmaformer av bare de leksikalske ordene. Hver kurve viser progressive verdier av entropi etter hvert som teksten "vokser" seg lengre.

Diagrammene viser altså kurver for håndtekst og tastetekst for enkeltelever, og i disse eksemplene er det klare differansetendenser hos 2 av 3 elever for lemmaformer, og for 3 av 3 for leksikalske lemmaformer. Hvis man skulle bruke dette i en statistisk test, får man imidlertid på grunn av den sterkt fallende tendensen det samme problemet og de samme utfordringene som for ubehandlede TTR-verdier, som forklart og diskutert i 10.2.

Forsøk med logaritmetransformering har ikke vist seg å ha den ønskede effekt på disse distribusjonene, og jeg har derfor i stedet gjort forsøk med en segmentbasert utregning, tilsvarende MOSTTR, altså med overlappende segmenter. Disse forsøkene er illustrert ved hjelp av tre tekster i figur 10-51 (for lemmaformer) og figur 10-52 (for leksikalske lemmaformer).

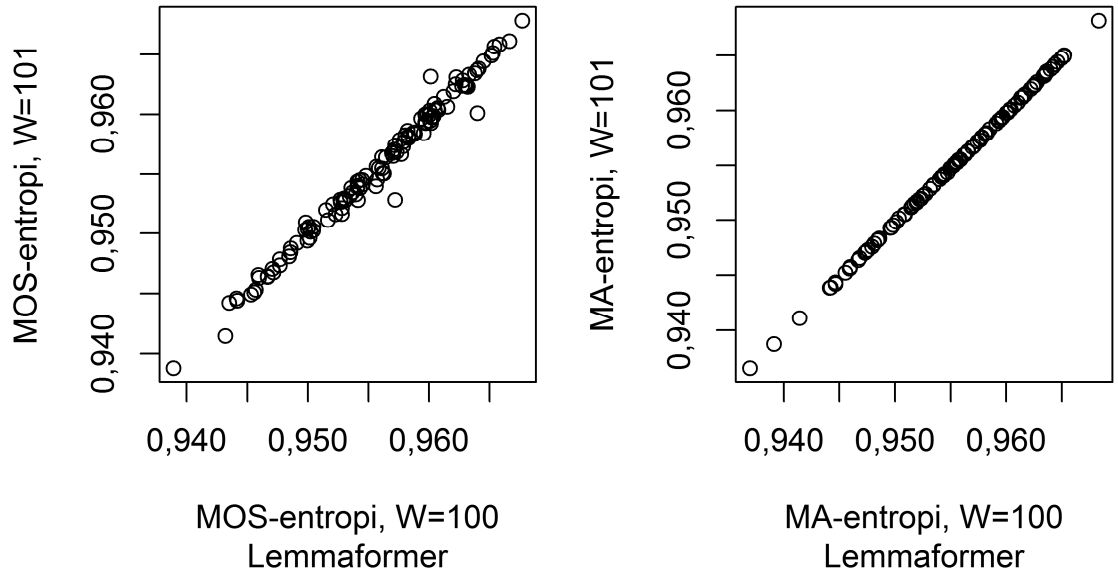


Figur 10-51: Progressive verdier for MOS-entropi basert på lemmaformer, W = 100.



Figur 10-52: Progressive verdier for MOS-entropi basert på leksikalske lemmaformer, W = 50.

Det mest umiddelbare inntrykket fra diagrammene er at de MOS-baserte entropimålene utviser den samme periodisiteten i progressive verdier som MOSTTR. Dette er ikke så overraskende, men svingningene virker kraftige for noen av tekstene, og det setter spørsmålstegn ved validiteten i dette MOS-baserte målet, tilsvarende problemene jeg pekte på for MOSTTR i 10.4.4. Et spredningsdiagram (det venstre diagrammet i figur 10-53 nedenfor) mellom MOS-entropi for lemmaformer med henholdsvis $W_1 = 100$ og $W_2 = 101$ illustrerer det samme problemet som figur 10-38 på side 215 avslørte for MOSTTR, nemlig at små endringer i W gir uforholdsmessig store utslag for enkelte tekster.



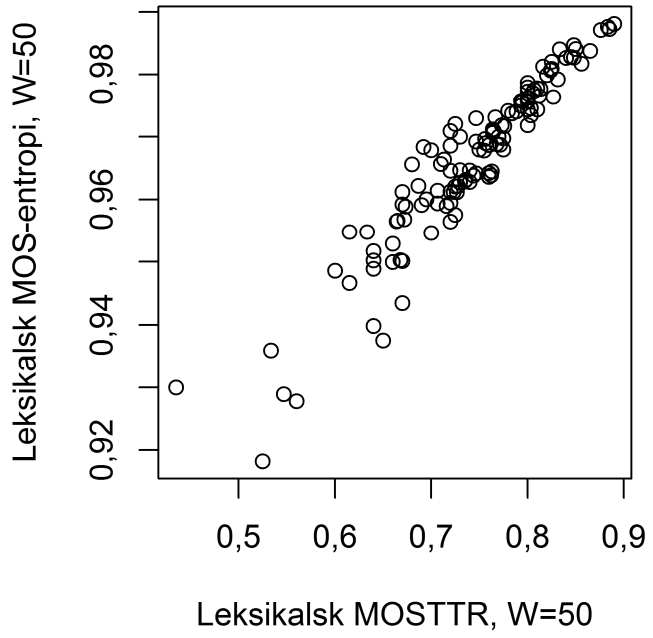
Figur 10-53: Korrelasjon mellom MOS-entropi (W=100) og MOS-entropi (W=101) til venstre og mellom MA-entropi (W=100) og MA-entropi (W=101) til høyre. Begge er basert på lemmaformer.

Dessuten ser vi at verdiene ligger ganske nær 1. Tabell 10-15 nedenfor viser at middelverdiene ligger i overkant av 0,95, med standardavvik på 0,006. Det er imidlertid ingen sterk tak-effekt, og fordelingen er normal ifølge Shapiro-Wilks normalitetstest, $W \approx 0{,}982$, $p \approx 0{,}10$, dog med en viss tendens til venstreskjevhet.

Tabell 10-15: Nøkkeltall for MOS-entropi (W=100) på lemmaformer

	middelverdi	median	sd	min	maks
Total	**0,956**	**0,956**	**0,006**	**0,939**	**0,968**
Hånd	0,955	0,956	0,006	0,939	0,967
Tast	0,956	0,957	0,006	0,943	0,968
Middels	0,955	0,956	0,006	0,939	0,967
Sterk	0,957	0,957	0,006	0,943	0,968
Gutt	0,958	0,957	0,005	0,946	0,968
Jente	0,954	0,954	0,006	0,939	0,965

På leksikalske ord korrelerer MOS-entropi$_{W=50}$ svært sterkt med MOSTTR-LL$_{W=50}$, R ≈ 0,95, N = 2x60). Dette tyder på at entropi *ikke* gir noen vesentlig gevinst over TTR, eller *ikke* måler noen annen eller mer avansert form for leksikalsk variasjon i teksten. Den sterke korrelasjonen indikerer sannsynligvis at entropimål krever større vindusbredde for å gi interessante resultater, og det er ikke mulig for leksikalske ord i dette materialet. Et spredningsdiagram (figur 10-54) viser nærmest sammenfall (selvfølgelig) mellom de to variablene når verdiene nærmer seg 1.



Figur 10-54: Korrelasjon mellom MOS-entropi (W=50) og MOSTTR (W=50) på leksikalske lemmaformer. R ≈ 0,95, N = 2x60.

Korrelasjonen mellom MOS-entropi$_{W=100}$ og MOSTTR$_{W=100}$ er vesentlig svakere når analysen omfatter alle lemmaformer, R ≈ 0.77. Dette skriver seg nok fra at funksjonsordene har mer interessante variasjonsdata for såpass smale vinduer, og det er et argument for å inkludere funksjonsordene i analyser basert på entropi. Det vil i alle fall være et verdifullt supplement til leksikalske analyser som er basert utelukkende på leksikalske ord, som MOSTTR-LL$_{W=50}$.

Sammenligning av de to diagrammene i figur 10-53 ovenfor viser at akkurat som for segmentbaserte TTR-mål kan den tilfeldige påvirkning av interaksjon mellom *W* og tekststruktur reduseres betraktelig ved å bruke *moving average*-beregninger i stedet for *mean overlapping segments*. I 10.4 argumenterte jeg for å benytte MOSTTR i stedet for MATTR, til tross for nettopp denne åpenbare ulempen knyttet til MOSTTR-verdiene. For entropi-analysene vil jeg heller legge vekt på fordelene ved *moving average* og bruke MA-entropi$_{W=100}$ på alle lemmaformer i de videre analysene.

10.5.2.1 Deskriptiv analyse

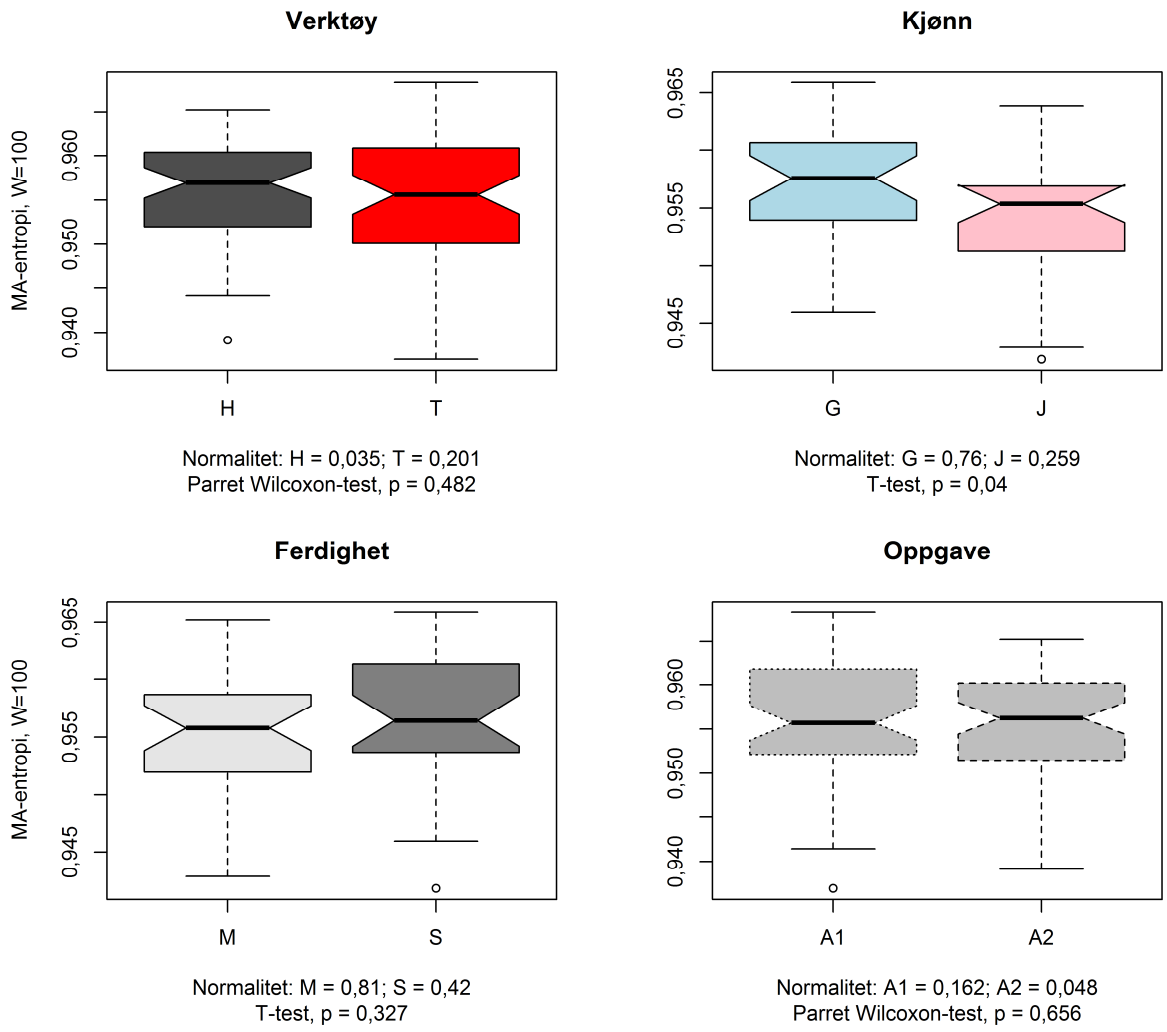
Nøkkeltallene for MA-entropi$_{W=100}$ i tabell 10-16 nedenfor viser at verdiene ligger på samme nivå som MOS-entropi$_{W=100}$, som ventet. Middelverdiene er i overkant av 0,95, med et standardavvik på 0,006. De ulike beregningsmåtene gir imidlertid en del variasjon i verdiene for de enkelte tekstene, og korrelasjonen mellom MA-entropi$_{W=100}$ og MOS-entropi$_{W=100}$ er bare $R \approx 0{,}91$, $N = 2x60$.

Tabell 10-16: Nøkkeltall for MA-entropi (W=100) på lemmaformer

	middelverdi	median	sd	min	maks
Total	**0,956**	**0,956**	**0,006**	**0,937**	**0,968**
Hånd	0,956	0,957	0,006	0,939	0,965
Tast	0,955	0,956	0,007	0,937	0,968
Middels	0,955	0,956	0,006	0,939	0,965
Sterk	0,956	0,957	0,007	0,937	0,968
Gutt	0,957	0,958	0,006	0,944	0,968
Jente	0,954	0,955	0,007	0,937	0,964

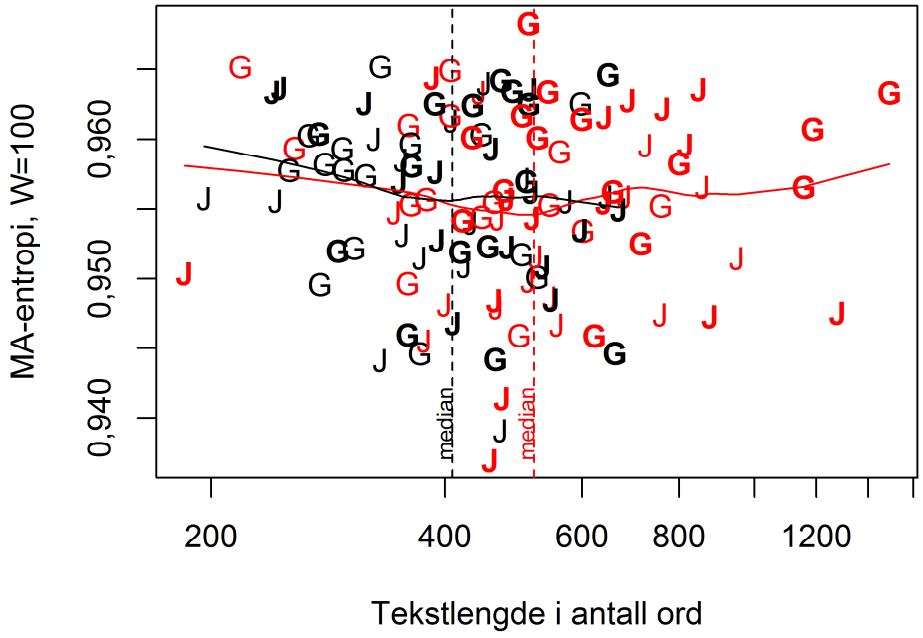
Distribusjonen er litt venstreskjev ($W \approx 0{,}973$, $p \approx 0{,}02$ med Shapiro-Wilks normalitetstest, $N = 2x60$), og selv om maksimumsverdiene ligger godt under 1, kan dette kanskje være et resultat av en takeffekt i en del av vinduene i noen av tekstene.

Figur 10-55 nedenfor viser at guttene har høyere verdier enn jentene, mens det er liten eller ingen effekt av skriveverktøy, ferdighet eller oppgave.



Figur 10-55: MA-entropi (W=100) for lemmaformer etter fire faktorer

Figur 10-56 nedenfor viser at det ikke er noen korrelasjon mellom variabelen og tekstlengde, noe som kanskje er litt overraskende i lys av korrelasjonen mellom tekstlengde og både $FSTTR_{W=400}$ og $MOSTTR_{W=100}$. Korrelasjonen mellom håndtekster og tastetekster er $R \approx 0{,}46$.



Figur 10-56: MA-entropi (W=100) og korrelasjon med tekstlengde

10.5.2.2 Variansanalyse

Variansanalysen er utført på den maksimale modellen med variabeldifferansen mellom tastetekster og håndtekster som responsvariabel og de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell som prediktorer, antall interaksjonsnivåer begrenset til 2:

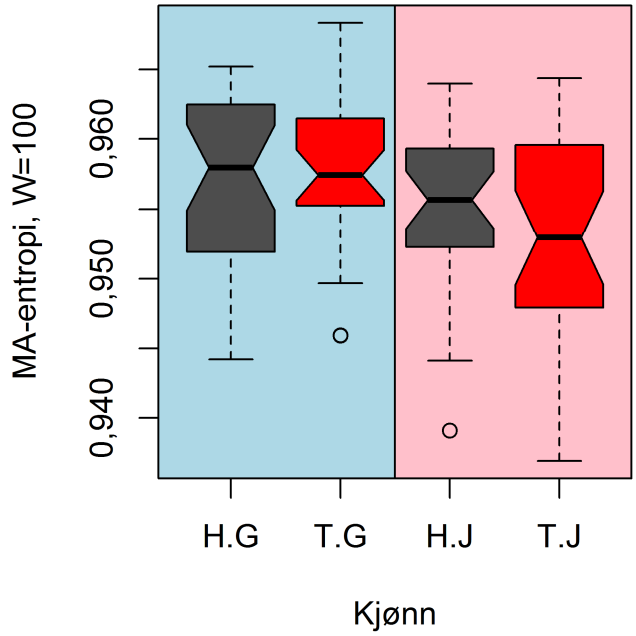
(125)
```
lm(lexD$MAentropi.lem.100~(kjønn+ferdighet+lengde+forskjell)^2)
```

Trinnvis reduksjon av anova-modellen resulterer i følgende minimale modell, med kjønn som eneste signifikante prediktor, $F \approx 4{,}32$, $p < 0{,}05$, $d \approx 0{,}55$:

(126)
```
lm(formula = lexD$MAentropi.lem.100 ~ kjønn)

             Df    Sum Sq   Mean Sq F value Pr(>F)
kjønn         1 0.0001802 1.802e-04   4.315 0.0422 *
Residuals    58 0.0024220 4.176e-05
---
Multiple R-squared:  0.06925,   Adjusted R-squared:  0.0532
F-statistic: 4.315 on 1 and 58 DF,  p-value: 0.04221
```

Tendensen er den samme som vi har sett i flere tidligere analyser, nemlig at jentene har en tendens til lavere verdier i tastetekstene, mens guttenes tekster ikke er påvirket av verktøyet. Se figur 10-57 nedenfor. `Gvlma` (se 7.2.2.4) viser at premissene for anova er oppfylt. (Se appendiks A4.)



Figur 10-57: MA-entropi på lemmaformer (W=100), resultat av anova-analyse

10.5.2.3 Diskusjon

Jeg pekte i 10.5.2.1 på at korrelasjonen mellom MOS-entropi$_{W=100}$ og MA-entropi$_{W=100}$ er overraskende lav, $R \approx 0{,}91$, sett i lys av at det er bare tekniske forskjeller mellom de to beregningsmåtene. Anova-analyse av MOS-entropi$_{W=100}$ gir også en annen minimal modell enn MA-entropi$_{W=100}$,[40] med interaksjon mellom tekstlengde og tekstlengdeforskjell som eneste signifikante faktor, $p < 0{,}01$. Dette viser at selv bare ganske små forskjeller mellom varianter av variabler kan ha konsekvenser for analyseresultatene og hvordan vi forstår tendensene i materialet. Det tilsier at vi bør tolke enkeltresultater med varsomhet.

Jeg tror entropi av ordtypefrekvenser er en variabel med potensial, men man må enten finne en måte å nøytralisere tekstlengdeeffekten på, eller ha lange nok tekster til at større vindusbredder kan brukes i segmentbaserte analyser. Kanskje er den særlig interessant med hensyn til å fange variasjon innenfor enkeltordklasser, eller for funksjonsord.

10.6 Oppsummering og diskusjon av leksikalske variabler

10.6.1 Oppsummering av resultater

Jeg har i dette kapitlet behandlet forskjellige mål for leksikalsk variasjon, og i det foregående kapitlet behandlet jeg forskjellige mål for informasjonell tetthet. Det er åpenbart at flere av

[40] Modellen som helhet er riktignok ikke signifikant, $F \approx 2{,}76$, $p \approx 0{,}0504$. Videre reduksjon av modellen gir ikke-signifikant null-modell som resultat. Se appendiks A2.

disse leksikalske målene er svært nært beslektet, og jeg har derfor ikke analysert alle like inngående med tanke på effekter knyttet til problemstillingen i avhandlingen. De leksikalske variablene jeg har valgt som de mest sentrale, og som blir brukt i prinsipalkomponentanalysen i kapittel 12, er de følgende:

- gjennomsnittlig ordlengde
- gjennomsnittlig ordlengde i leksikalske ord
- leksikalsk tetthet
- log-$TTR_{1,3}$ (global TTR)
- MOSTTR-$LL_{W=50}$ (lokal TTR)

Også disse fem variablene er det en del samvariasjon mellom, slik det går fram av tabell 10-17 nedenfor.

Tabell 10-17: Pearsons korrelasjonskoeffisienter mellom fem sentrale leksikalske variabler. Ikke alle variablene er normalfordelt, og det er brukt 120 observasjoner i analysene, hvorav bare 60 er uavhengige.

	Ordlengde	**Leksikalsk ordlengde**	**Leksikalsk tetthet**	**Global TTR**	**Lokal TTR**	**Tekst-lengde**
Ordlengde	–	0,85	0,68	0,54	0,34	–0,07
Leks. ordlengde	0,85	–	0,33	0,57	0,53	0,02
Leks. tetthet	0,68	0,33	–	0,30	0,04	–0,19
Global TTR	0,54	0,57	0,30	–	0,85	0,29
Lokal TTR	0,34	0,53	0,04	0,85	–	0,25
Tekstlengde	–0,07	0,02	–0,19	0,29	0,25	–

Først og fremst, og ikke overraskende, er det sterk korrelasjon mellom gjennomsnittlig ordlengde og gjennomsnittlig ordlengde for leksikalske ord. Man kan lure på om denne korrelasjonen er så sterk at de to variablene i realiteten er uttrykk for samme tekstlige egenskap, og at de dermed overflødiggjør hverandre. Samtidig har disse to også noen forskjellige egenskaper; den siste varianten er mye mindre avhengig av leksikalsk tetthet og samvarierer vesentlig mer med MOSTTR-$LL_{W=50}$.

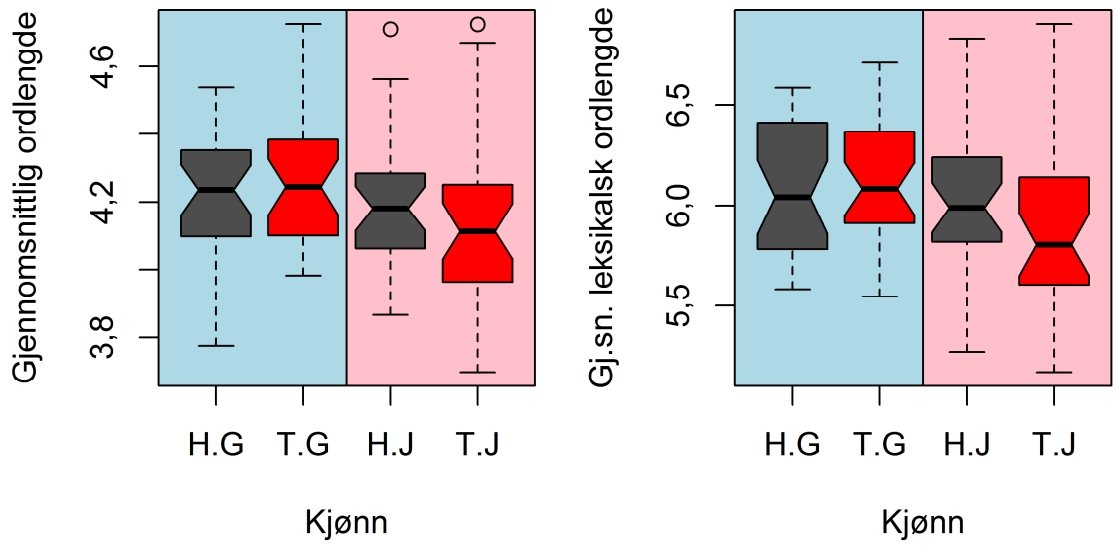
Dernest er det verdt å legge merke til den sterke korrelasjonen mellom log-TTR og MOSTTR-LL, selv om MOSTTR-LL i motsetning til log-TTR er lokal og basert bare på leksikalske ord. Det illustrerer at en stor del av det som bidrar til den globale log-TTR, er ganske lokal variasjon. Derimot er det ingen sammenheng mellom MOSTTR-LL og leksikalsk tetthet. Jeg fjernet de grammatiske ordene fra MOSTTR-analysen nettopp med tanke på å nøytralisere den leksikalske tettheten som bidragsyter til variabelen, men en viss sammenheng mellom leksikalsk tetthet og leksikalsk variasjon kunne man likevel ha ventet, på linje med at det er korrelasjon mellom leksikalsk ordlengde og leksikalsk tetthet.

Når det gjelder sammenhengen mellom type/eksemplar-forholdstall og tekstlengde, så er altså FSTTR, MSTTR og MOSTTR konstruert slik at den matematiske sammenhengen med tekstlengde nøytraliseres ved at verdiene regnes ut på grunnlag av like tekstsegmentlengder. Log-TTR er derimot konstruert slik at det ikke skal være noen korrelasjon med tekstlengde for det aktuelle tekstutvalget, og deretter justert slik at sammenhengen med tekstlengde

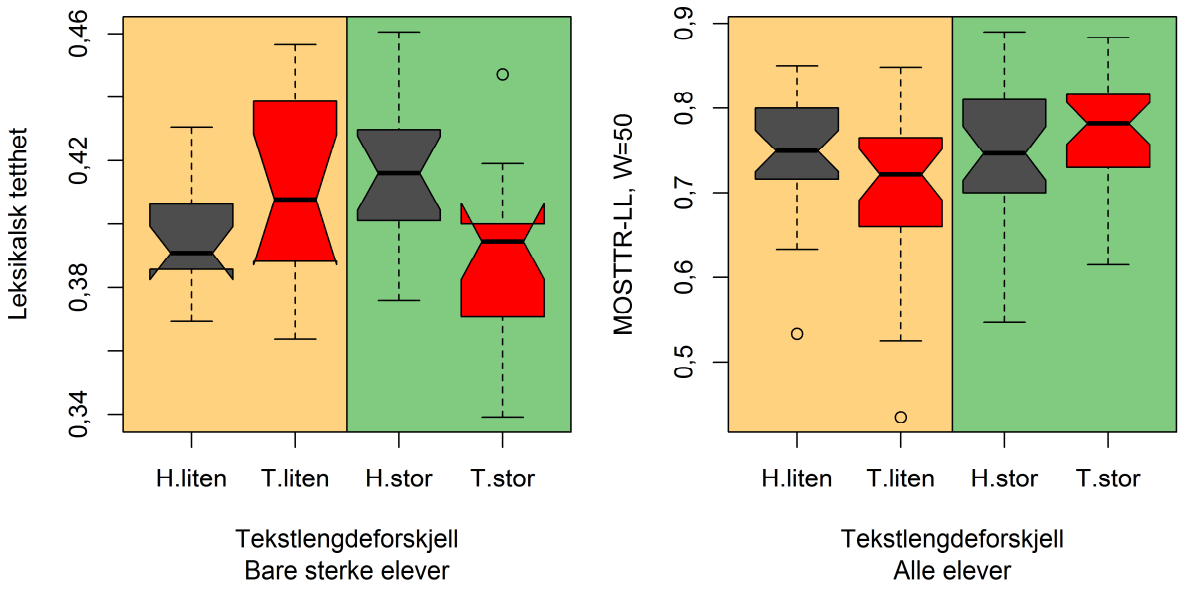
ligner på den for FSTTR og MOSTTR. Alle disse målene korrelerer positivt med tekstlengde; verdiene tenderer altså til å være høyere i lengre tekster. Siden dette ikke skyldes matematiske egenskaper forbundet med tekstlengden, må det for FSTTR, MSTTR og MOSTTR skyldes globale egenskaper ved tekstene. Vi kan altså anta at lengre tekster i gjennomsnitt har større leksikalsk variasjon i dette materialet, også større lokal variasjon.

Dette kan forklares i et elevperspektiv eller i et tekstperspektiv. Med utgangspunkt i elevene er det nærliggende å tenke seg at forklaringen ligger i at sterke elever skriver lengre, og at de også har høyere leksikalsk variasjon. Med utgangspunkt i tekstene kan sammenhengen forklares ut fra at økt lengde er et resultat av at tekstene inneholder flere momenter, og at variasjon i momenter fører til variasjon i ordbruk. Disse to perspektivene utelukker selvfølgelig ikke hverandre, siden man også kan regne med at det gjerne er de sterke elevene som har størst innholdsrikdom i tekstene. Man må likevel være forsiktig med å trekke slutninger om at leksikalsk variasjon nettopp er et uttrykk for kvalitet. Stor leksikalsk variasjon kan også være et uttrykk for svak koherens.

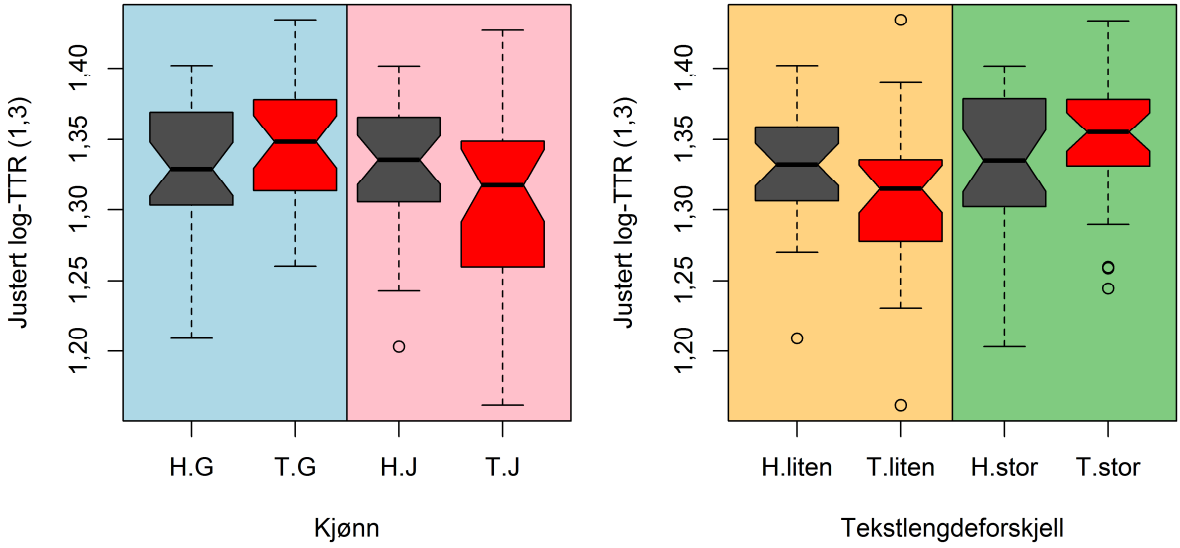
Figur 10-58, figur 10-59 og figur 10-60 nedenfor oppsummerer de signifikante resultatene fra anova-modelleringen av disse fem variablene, og diagrammene avdekker flere beslektede egenskaper.



Figur 10-58: Ordlengde til venstre, og leksikalsk ordlengde til høyre. Jenter har tydelig lavere verdier i tastetekstene for begge variabler.



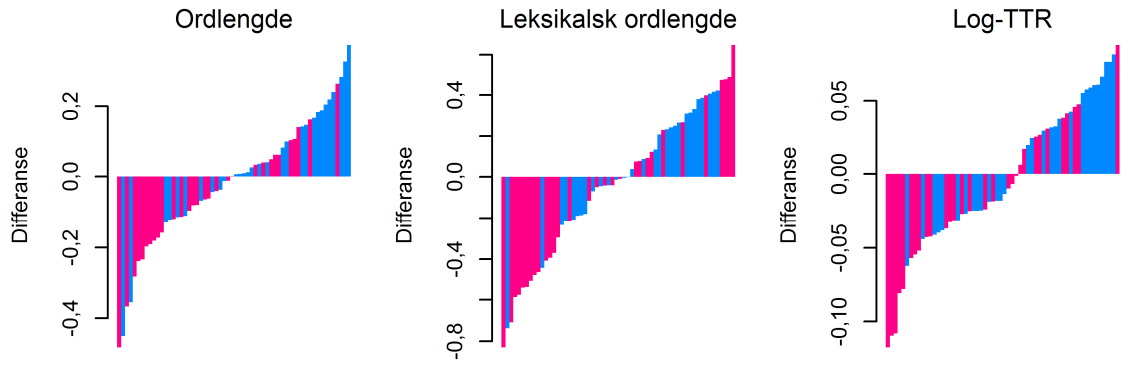
Figur 10-59: Leksikalsk tetthet til venstre (bare sterke elever), og MOSTTR-LL$_{W=50}$ til høyre (alle elever). Sterke elever med liten forskjell i tekstlengde har høyere leksikalsk tetthet i tastetekstene, mens sterke elever med stor tekstlengdeforskjell har lavere leksikalsk tetthet i tastetekstene. Elever med liten tekstlengdeforskjell har lavere MOSTTR i tastetekstene, mens elever med stor tekstlengdeforskjell har høyere MOSTTR i tastetekstene.



Figur 10-60: Log-TTR justert med faktor 1,3. Gutter har noe høyere log-TTR i tastetekstene, mens jentene har lavere log-TTR i tastetekstene (til venstre). Elever med liten tekstlengdeforskjell har lavere log-TTR i tastetekstene, mens elever med stor tekstlengdeforskjell har høyere log-TTR i tastetekstene (til høyre).

Begge ordlengdemålene og log-TTR har omtrent den samme distribusjonen for kjønn; jentene har lavere verdier i tastetekstene, mens guttene har høyere verdier i tastetekstene. Effekten for guttene er noe svakere enn for jentene og kommer ikke godt fram i alle boksdiagrammene. Det samme gjelder for log-TTR, og her har guttene tydelig høyere verdier i tastetekstene. I de andre tendensene vi ser, er det tekstlengdeforskjell som påvirker. Både log-TTR og MOSTTR er kjennetegnet ved høyere tasteverdier for elever som skriver mye lengre på tastatur, og lavere tasteverdier for de andre elevene. For leksikalsk tetthet er

det en forskjell mellom middels og sterke elever; for middels elever har verktøyet ingen effekt på den leksikalske tettheten, mens blant de sterke elevene har elever med liten tekstlengdeforskjell høyere leksikalsk tetthet i tastetekstene, mens elever med stor tekstlengdeforskjell har lavere leksikalsk tetthet i tastetekstene, altså en helt annen type effekt enn for de to TTR-variablene. Hovedsakelig har ferdighet og total tekstlengde ingen effekt på de leksikalske variablene.



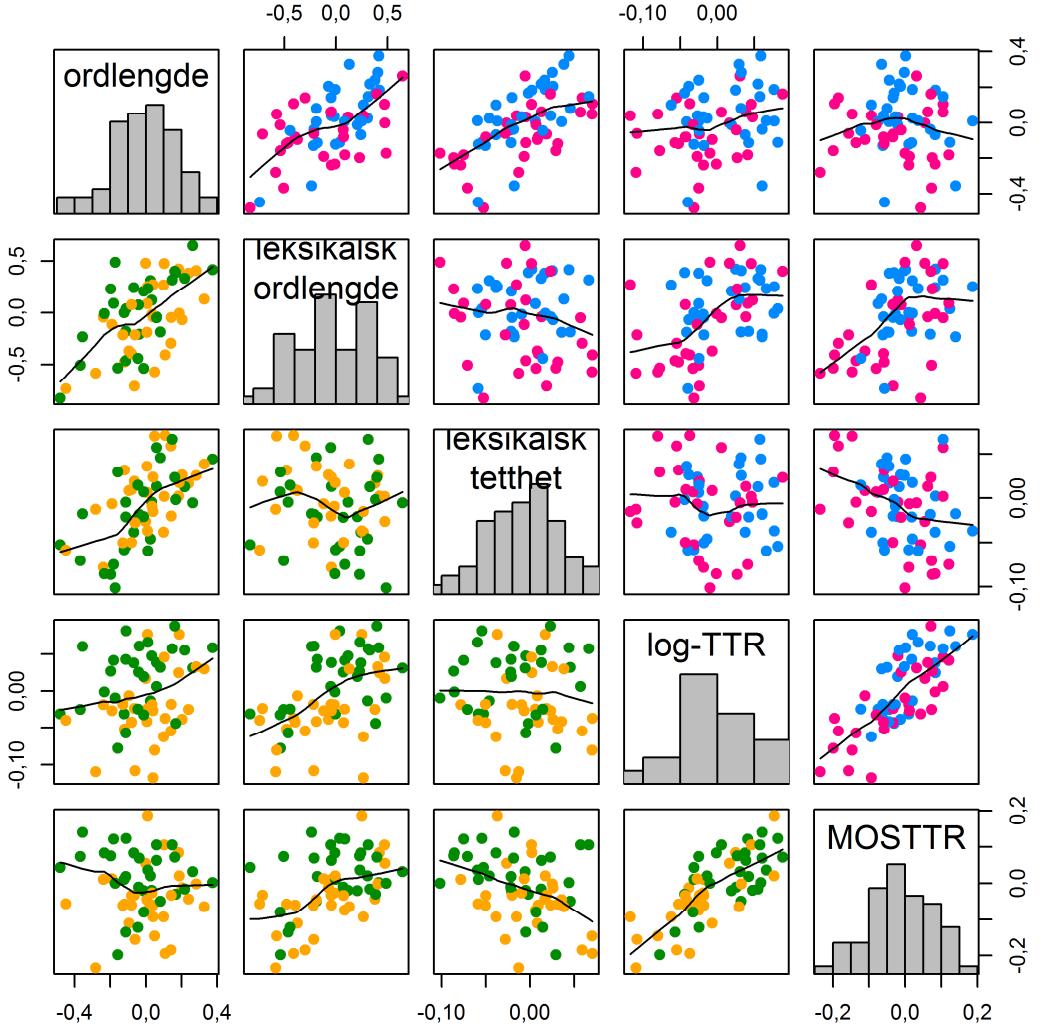
Figur 10-61: Ordlengde, leksikalsk ordlengde og log-TTR. Skreddiagrammer for differanseverdier som viser effekten av kjønn.

I skreddiagrammer (figur 10-61) kommer kjønnseffekten tydelig fram ved at de to endene i diagrammene er dominert av hvert sitt kjønn, men spredningsdiagrammene i figur 10-62 nedenfor viser at det ikke er de samme elevene som har lave verdier i alle variablene, slik man kanskje kunne tro. Det er altså ikke slik at alle kjønnsforskjellene skyldes én og den samme subgruppen av jenter og én og den samme subgruppen av gutter. Derimot ser det ut til å være en mer generell kjønnseffekt som kommer til uttrykk på ulike måter i ulike individer.

Krysskorrelasjonsdiagrammet og korrelasjonstabellen i figur 10-62 og tabell 10-18 nedenfor viser korrelasjonene mellom elevenes differanseverdier for de fem leksikalske variablene, altså i motsetning til korrelasjonene mellom tekstverdiene, som vi har sett på tidligere. Tabellen viser at de sterke korrelasjonene mellom ordlengdevariablene og mellom variasjonsvariablene ikke er like sterke når vi ser på differanseverdiene. Dette tyder på at det er forskjell mellom variabelvariantene som gjør det verdt å bruke begge. Korrelasjonen mellom de to variasjonsvariablene er riktignok fortsatt ganske sterk, $R \approx 0{,}75$, men ikke ekstremt sterk, og jeg tror det er fornuftig å beholde to variabler for henholdsvis global og lokal leksikalsk variasjon. Den middels sterke korrelasjonen mellom begge de to variablene og forskjell i tekstlengde svarer godt til resultatet fra variansanalysen.

Verdt å merke seg er at selv om variansanalysene av ordlengde og leksikalsk tetthet gir helt ulikt resultat, er det så sterk korrelasjon mellom deres differanseverdier som $R \approx 0{,}61$. Denne korrelasjonen er ventet og ikke overraskende ut fra diskusjonen om sammenhengen mellom ordlengde og leksikalsk tetthet, og den illustrerer at variansanalysene ikke nødvendigvis fanger opp alle relevante egenskaper ved variablene. Prinsipalkomponentanalysen i kapittel 12 (se f.eks. tabell 12-7) viser imidlertid sammenhengen mellom dem ved at de er de to

viktigste variablene i dimensjon 2, noe som også illustrerer verdien av en multivariat analyse som ser helheten i systemet av variabler.



Figur 10-62: Korrelasjoner mellom differanseverdier for de fem valgte leksikalske variablene

Tabell 10-18: Pearsons korrelasjonskoeffisienter mellom differanseverdier for de fem valgte leksikalske variablene pluss tekstlengde

Differanse	**Ordlengde**	**Leksikalsk ordlengde**	**Leksikalsk tetthet**	**log-TTR**	**MOSTTR**	**Tekst-lengde**
Ordlengde	–	0,59	0,61	0,26	–0,05	–0,23
Leks. ordlengde	0,59	–	–0,12	0,49	0,47	0,15
Leks. tetthet	0,61	–0,12	–	–0,06	–0,37	–0,38
log-TTR	0,26	0,49	–0,06	–	0,75	0,45
MOSTTR	–0,05	0,47	–0,37	0,75	–	0,44
Tekstlengde	–0,23	0,15	–0,38	0,45	0,44	–

Når det gjelder korrelasjonene med tekstlengde, må vi være oppmerksom på at den i realiteten i hvert fall delvis kan være en effekt av oppgave, ettersom det er så stor forskjell i tekstlengde mellom oppgavene. Med den valgte designen er det imidlertid vanskelig å avgjøre hvorvidt eller i hvilken grad oppgaven spiller inn.

Hvis tekstlengdeeffekten imidlertid faktisk er en reell effekt av tekstlengde, så ser det altså ut til at de som skriver like langt med begge verktøy, skriver med lavere variasjon på tastatur, mens de som skriver lengre på tastatur, skriver med høyere variasjon på tastatur. Dette strider i utgangspunktet mot hypotesen om at raskere produksjon fører til mer spontan språkbruk, men effekten kan ha flere alternative forklaringer. En mulig forklaring er at de som blir inspirert til å skrive lengre, også greier å skrive mer variert, kanskje fordi de har mer å skrive om. En annen forklaring kan være at de som skriver lengre, gjør det fordi de har raskere skrivehastighet, og at de dermed også greier å utnytte denne ferdigheten til å lage en mer variert tekst, kanskje fordi de likevel får mer tid til redigering eller formulering. Et siste alternativ til forklaring er at en del av de som skriver langt, bare bruker tastaturferdighetene til å produsere mye tekst med mindre indre sammenheng og dermed høyere variasjon, slik vi så i eksempel (118) på side 193 fra en av de lengste tekstene. Dette kan kanskje tolkes som en form for spontanitet i språkbruken, men i en litt annen forstand enn den som lå til grunn for hypotesen. Disse tre mulige forklaringene utelukker hverandre selvfølgelig ikke, og de kan gjelde ulike elever i ulik grad.

10.6.2 Diskusjon av validitet

Som Jarvis (2013) er inne på, har mange diskusjoner om leksikalske mål dreid seg om verifiserbarhet. Mye av drøftingen i avsnittene over har også mest fokus på verifiserbarhet og mindre på validitet. Jeg har lagt vekt på å utvikle målemetoder som i størst mulig grad måler variasjon basert på hele teksten, og som i minst mulig grad blir påvirket av tilfeldig interferens mellom *W* og *N* eller mellom *W* og variasjonsvariasjon i teksten, eller av hvordan forholdet mellom antall typer og antall eksemplarer utvikler seg matematisk gjennom en tekst.

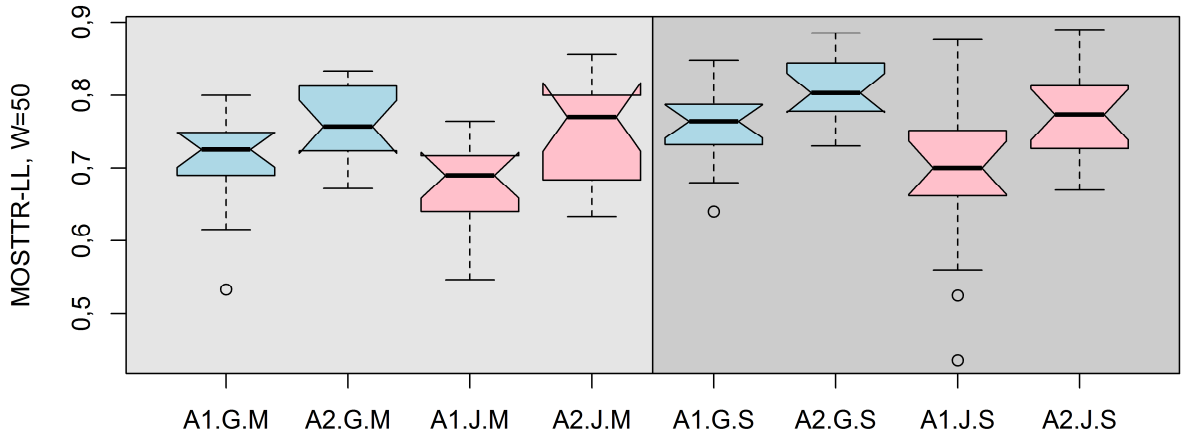
Jeg stiller imidlertid i liten grad spørsmålet om antall unike ord dividert på antall løpeord er et valid mål på ordvariasjon i seg selv, altså om det forteller oss det vi ønsker å vite om ordvariasjonen i en tekst. Det er lett å la seg villede av at TTR-målene er uttrykt som forholdstall eller desimaltall, og man kan lett få en følelse av at disse forholdstallene uttrykker mer enn de faktisk gjør. TTR-målene som FSTTR, MATTR eller MOSTTR representerer faktisk ikke noe mer sofistikert enn *antall* ulike ord i et tekstsegment av en viss lengde. Det er minst to måter dette er et utilfredsstillende mål på.

For det første forteller et tall for antall ordtyper lite om den tallmessige variasjonen som finnes blant disse ordene. Som Jarvis er inne på, kan kanskje et entropibasert variasjonsmål være bedre egnet til å gjenspeile den reelle variasjonen mellom ordtyper og ordeksemplarer, og jeg har eksperimentert med entropi (i 10.5.2) og hapax legomena (i 10.5.1), men uten å komme fram til en variabel med tilstrekkelig solide egenskaper.

For det andre fanger tallmessige beregninger av repetisjon av denne typen i liten grad hvor leksikalsk rik, spesifikk, nyansert eller sofistikert teksten er (Jarvis, 2013), tekstlige egenskaper som vi nok er mer interessert i enn grad av repetisjon av ordformer eller leksemer. Til en viss grad kan man kanskje avdekke slike egenskaper når man kombinerer repetisjonsbaserte analyser med leksikalsk tetthet, ordlengde og analyser basert på ordenes frekvens i språket generelt. Men generelt må vi nok regne med å måtte kombinere de kvantitative analysene med mer kvalitative tekstanalyser for å kunne avdekke alle de egenskapene vi er interessert i.

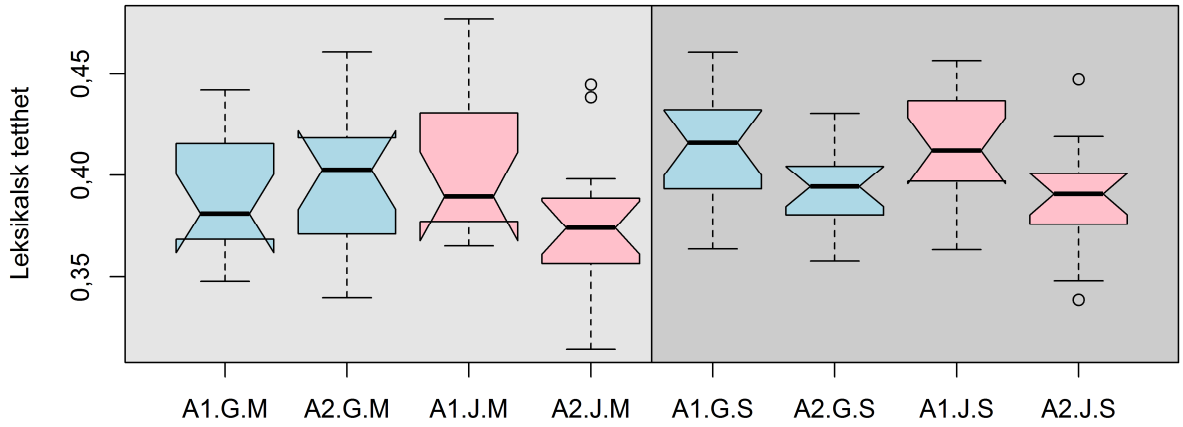
Det er likevel verdt å merke seg den sterke korrelasjonen mellom gjennomsnittlig ordlengde og både leksikalsk tetthet og leksikalsk variasjon. Dette svært enkle, mekaniske målet reflekterer trolig flere ulike relevante tekstlige egenskaper og kan brukes i hvert fall som en rask tilnærming til en leksikalsk analyse av en tekstsamling.

Til slutt vil jeg peke på et annet faktum ved TTR-baserte mål som ikke direkte angår skriveverktøyet. Lærerne rapporterte at de syntes "Ungdomsfylla"-oppgaven fungerte bedre enn "Bøker eller data"-oppgaven. Elevene fikk i større grad til å diskutere synspunktet i oppgaven. Figur 10-63 nedenfor viser også at alle utvalgssegmenter av kjønn og ferdighet har høyere MOSTTR-LL i "Ungdomsfylla"-tekstene. Dette *kan* indikere at MOSTTR-LL faktisk *er* et mål som korrelerer med tekstkvalitet, og at det kanskje er et symptom på momentrikdom eller nyansering i tekst.



Figur 10-63: MOSTTR-LL$_{W=50}$ og sammenheng med oppgave, kjønn og skriveferdighet. A1 = "Bøker eller data", A2 = "Ungdomsfylla", G = Gutter, J = Jenter, M = Middels elever, S = Sterke elever.

Når det gjelder leksikalsk *tetthet*, korrelerer den svakt med leksikalsk variasjon i form av log-TTR, men ikke med MOSTTR-LL. Figur 10-64 nedenfor viser at interaksjonen mellom oppgave, kjønn og ferdighet er en helt annen; hovedtendensen er at "Ungdomsfylla"-tekstene har lavere leksikalsk tetthet, bortsett fra for de middels sterke guttene, der effekten er svak men motsatt. Hvis vi holder disse guttene utenfor, kan dette kanskje tyde på at engasjert skriving fører til lavere leksikalsk tetthet for disse elevene.



Figur 10-64: Leksikalsk tetthet og sammenheng med oppgave, kjønn og skriveferdighet. A1 = "Bøker eller data", A2 = "Ungdomsfylla", G = Gutter, J = Jenter, M = Middels elever, S = Sterke elever.

11 Syntaks

I kapittel 9 og 10 diskuterer og analyserer jeg ulike leksikalske variabler knyttet til tetthet og variasjon. Kapitlene viser at flere tekstlige egenskaper gjenspeiles av leksikalske variabler. Men som Halliday (f.eks. 1989) hevder, Biber (1988) viser, og jeg er inne på i kapittel 2, henger leksikalske variabler tett sammen med syntaks, og syntaktiske og leksikalske variabler påvirker hverandre og utfyller hverandre. En analyse av leksikalske trekk blir derfor stående litt ustøtt om den ikke suppleres med syntaktiske trekk, og jeg analyserer og drøfter derfor i dette kapitlet syntaktiske variabler relatert til ulike former for syntaktisk kompleksitet. I neste kapittel vil jeg se nærmere på samspillet mellom de leksikalske og de syntaktiske variablene. Dette samspillet er det som motiverer begrepet *leksikosyntaktiske trekk*, som er del av tittelen til denne avhandlingen.

Dette kapitlet begynner med ulike blikk på syntagmelengde (11.1), før de to neste utdyper de to mekanismene som påvirker syntagmelengden, nemlig antall ledd (11.2) og leddenes lengde (11.3). Til slutt i kapitlet kommer en oppsummering og diskusjon av funnene som er presentert i kapitlet (11.4).

Den metodiske tilnærmingen til utvalg av syntaktiske variabler er noe annerledes enn for de leksikalske variablene. Jeg tar utgangspunkt i t-enheten som et grunnleggende syntaktisk segment, drøfter ulike prinsipielle måter t-enhetens kompleksitet kan påvirkes, og lar denne drøftingen ligge til grunn for valg av en håndfull variabler. Drøfting av de enkelte variablenes tallmessige egenskaper som grunnlag for å evaluere deres validitet får dermed en mindre fremtredende rolle enn for de leksikalske variablene.

Jeg tar som utgangspunkt en enkel analyse av gjennomsnittlig t-enhetslengde i tekstene, målt i antall ord. Analysen i 11.1.1 nedenfor viser at tastetekstene har lengre t-enheter enn håndtekstene. Det er prinsipielt to måter å forlenge t-enheter på; antall ledd kan økes, og leddenes lengde kan økes (Hunt, 1965, s. 36). Leddenes lengde kan dessuten økes på to måter, med og uten klausal underordning; subklaususer kan fylle eller inngå i leddene, eller leddene kan fylles av lengre fraser uten subklaususer. Når det gjelder antall ledd, kan det også potensielt være fruktbart å skille mellom obligatoriske og adjungerte ledd; visse hovedverb deler ut flere roller enn andre, så antall ledd kan økes gjennom å velge hovedverb med mer kompleks argumentstruktur eller gjennom å legge til adjungerte ledd. I analysene i denne avhandlingen er fokuset på adjungerte ledd og ikke på bruk av to- eller trevalente hovedverb.

Dette kapitlet utforsker de to fasettene av syntaktisk kompleksitet som er skissert i forrige avsnitt: antall ledd og leddenes lengde. For hver fasett har jeg valgt et lite antall operasjonaliseringer. Operasjonaliseringene kan i liten grad sies å representere de aktuelle egenskapene direkte, men de måler språklige egenskaper som er relatert til de aktuelle egenskapene. Som en inngang til analysen av de to fasettene har jeg analysert tre variabler som er mer overordnet knyttet til syntagmelengde, og som vi kan anta vil korrelere med variablene som er knyttet til fasettene.

- Neksussyntagmelengde
    - Gjennomsnittlig antall ord i t-enheten
    - Gjennomsnittlig antall ord i klaususen
    - Frekvens av korte subklaususer
- Antall ledd
    - Preposisjonsfraser
    - Adverbiale subklaususer
- Leddenes lengde
    - Subklaususfrekvens
    - Forfeltets lengde
    - Attributive adjektiver

11.1 Neksussyntagmenes lengde

På samme måte som tekstenes lengde kan måles på flere måter (8.4.1), er det også flere alternative mål for syntagmelengde. De mest aktuelle er antall ord og antall leksikalske ord.

Dersom det er riktig som Halliday (1987, s. 62) hevder, at muntlige ytringer blir lengre enn skriftlige først og fremst fordi antall grammatiske ord er høyere, mens antall leksikalske ord er noenlunde konstant, er antall leksikalske ord kanskje et mål for syntagmelengde som i større grad fanger syntagmets "størrelse", innholdsmessige omfang eller kompleksitet.

Ett av argumentene for ikke å bruke antall leksikalske ord som mål for tekstlengde (8.4.1), er at korrelasjonen mellom antall ord og antall leksikalske ord i tekstene er så sterk ($R \approx 0{,}98$, $N = 2x60$) at valget trolig har liten praktisk betydning. I så fall er det mest fornuftig å velge det målet som det hefter minst teoretisk usikkerhet ved. Som mål på gjennomsnittlig lengde av syntagmer som t-enhet og klausus er imidlertid ikke korrelasjonen like sterk, $R \approx 0{,}88$ for t-enheter, $R \approx 0{,}86$ for klaususer, og det vil dermed trolig ha konsekvenser for konklusjonene om man velger det ene eller det andre. At korrelasjonen er svakere for disse syntagmetypene enn for tekstene som helhet, viser også at det finnes relasjoner mellom leksikalsk tetthet og syntagmelengder i materialet, slik Halliday peker på. Siden mengden av leksikalske ord ikke har noen klar teoretisk avgrensning (se diskusjonen i 9.2.1), er det tryggest å ta utgangspunkt i den enheten som det hefter minst teoretisk usikkerhet ved, nemlig ordet.

Søkene er gjort i det CG3-taggede korpuset (se kapittel 6.4.2), og tallene omfatter dermed noen få flerords leksemer, totalt 196.

11.1.1 T-enhetslengde

Jeg tar utgangspunkt i lengden til det mest grunnleggende segmentet i korpuset, nemlig t-enheten.

11.1.1.1 Hypotese

Jeg kjenner ikke til noe solid teoretisk fundament å forme hypoteser på når det gjelder t-enhetslengde. Halliday (1979, s. 49; 1989, s. 61-91) peker på at setningers kompleksitet bygges på ulike måter i muntlig og skriftlig språk, men han ser ikke ut til å basere dette på noe egentlig empirisk materiale, og han ser heller ikke ut til å hevde at den påstått høyere klaususkompleksiteten i muntlig språk fører til lengre *clause complexes* enn i skriftlig. Chafe & Danielewicz (1987, s. 96) finner at intonasjonsenheter er lengre i skriftlig språk enn i muntlig, men intonasjonsenheter kan ikke direkte sammenlignes med t-enheter. Hunt (1965, s. 23) finner, ikke overraskende, at t-enhetslengder øker med økende skrivekyndighet (se også Hudson (2009, s. 349)), men resultatene hans er ikke direkte relevante for forskningsspørsmålet i denne avhandlingen.

Jeg har dermed ingen entydig hypotese for t-enhetslengde i elevtekstmaterialet, men variabelen kan sees som et grunnleggende og uteoretisk mål for kompleksitet. Med en slik enkel tilnærming kan vi se på lengre t-enheter som mer komplekse enn kortere t-enheter. T-enhetens lengde kan imidlertid økes både ved mer klausal underordning og ved lengre eller flere ledd uten klausal underordning, og jeg knytter i hypotesen min disse to måtene til hvert sitt skriveverktøy. Jeg analyserer derfor t-enhetslengde først og fremst som et utgangspunkt for de mer nyanserte analysene av trekk knyttet til ulike former for syntaktisk kompleksitet.

11.1.1.2 Korpussøk

Jeg tar altså utgangspunkt i antall ord som mål på syntagmelengde. Overskrifter, signaturer og det jeg kaller ekte fragmenter (se 4.1.3), regnes ikke med. Det jeg kaller finitte fragmenter, altså fragmenter som inneholder finitte verbaler og er tagget som `<t-unit type="frag">` i korpuset, er derimot regnet med. Dette gjør det enkelt å hente ut de relevante tallene fra korpuset slik det er tagget:

```
(127) [lemma!="\$.*" & !<> & path=("t-unit")]
(128) <t-unit>
(129) syn$TEL <- lex$n.ordeks / syn$TE
```

(127) henter ut alle ord som ikke er skilletegn, som ikke er strukturelle attributter, og som står innenfor en t-enhet, inkludert finitte fragmenter. (128) henter ut alle t-enheter, inkludert t-enhetene av typen finitte fragmenter, men ikke ekte fragmenter. (129) regner ut gjennomsnittlig t-enhetslengde for hver tekst ved å dividere antall ord fra (127) på antall t-enheter fra (128).

11.1.1.3 Deskriptiv analyse

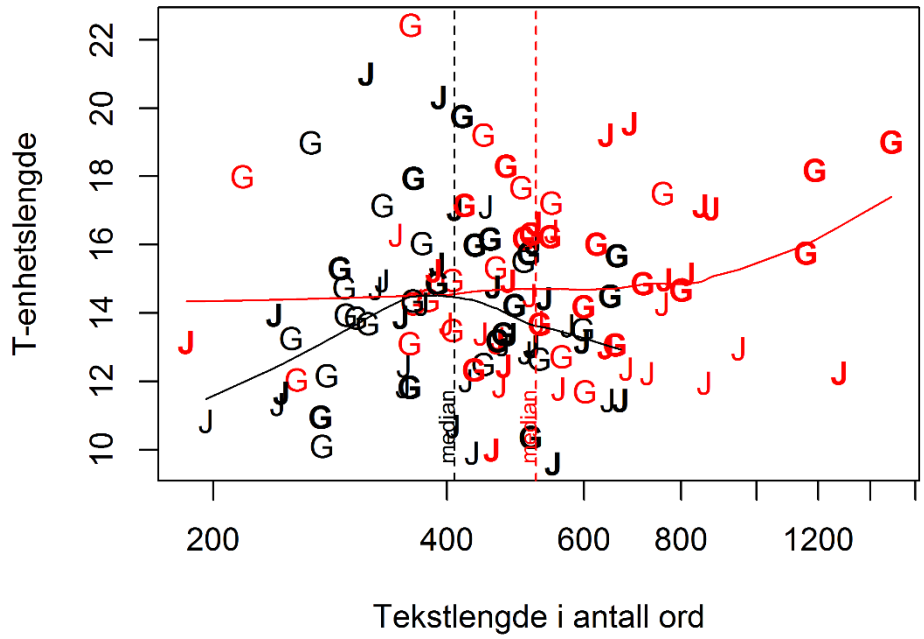
Tabell 11-1 viser nøkkeltallene for tekstenes gjennomsnittlige t-enhetslengde.

Tabell 11-1: Nøkkeltall for t-enhetslengde

	middelverdi	median	sd	min	maks
Total	**14,51**	**14,27**	**2,52**	**9,60**	**22,44**

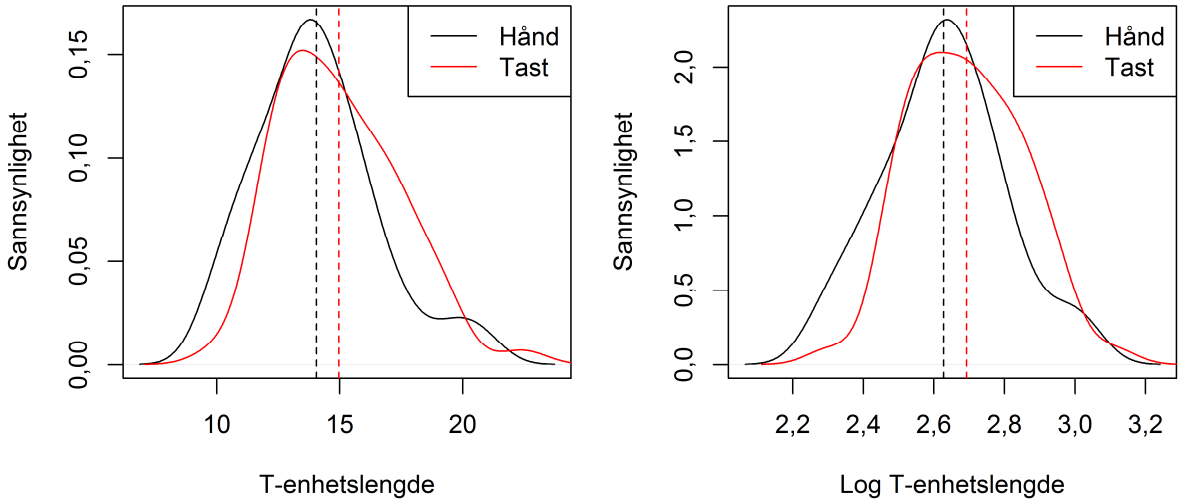
	middelverdi	median	sd	min	maks
Hånd	14,05	13,87	2,52	9,60	21,00
Tast	14,96	14,76	2,45	10,00	22,44
Middels	14,02	13,66	2,40	9,91	22,44
Sterk	14,99	14,91	2,56	9,60	21,00
Gutt	15,05	14,81	2,46	10,11	22,44
Jente	**13,96**	**13,51**	**2,47**	**9,60**	**21,00**

Tabellen viser at middelverdiene ligger mellom 14 og 15 ord per t-enhet, men tekstene med lengst t-enheter har *gjennomsnittlig* mer enn dobbelt så lange t-enheter som de med kortest. Det er interessant å vite om denne spredningen korrelerer med tekstlengde, men figur 11-1 viser at dette ikke er entydig. Den eneste effekten det synes rimelig å regne med, er at t-enhetslengde korrelerer positivt med tekstlengde for korte håndtekster, altså at de som skriver korte tekster for hånd, også skriver kortere t-enheter. De som skriver tilsvarende korte tekster på tastatur, er så få at det er umulig å konkludere enten om korrelasjon eller mangel på korrelasjon.



Figur 11-1: T-enhetslengde og tekstlengde

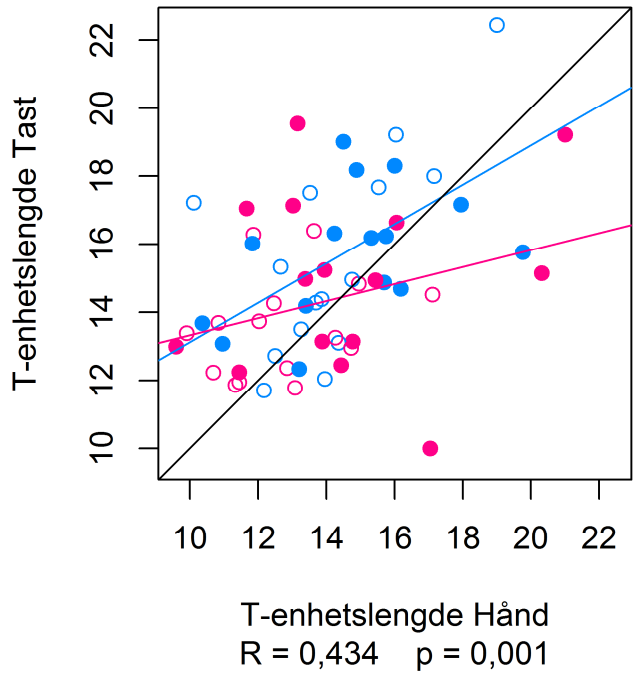
En nærliggende tolkning av denne tendensen er at elever som skriver kort, har "lite å skrive om", og at de på grunn av manglende engasjement eller få ideer produserer korte, lite komplekse t-enheter med relativt lite innhold.



Figur 11-2: Normalitet for t-enhetslengde. Til venstre utransformerte verdier; til høyre logaritmetransformerte verdier. De oppgitte normalitetsverdiene er p-verdier fra Shapiro-Wilks normalitetstest for henholdsvis hele utvalget, håndtekstene og tastetekstene.

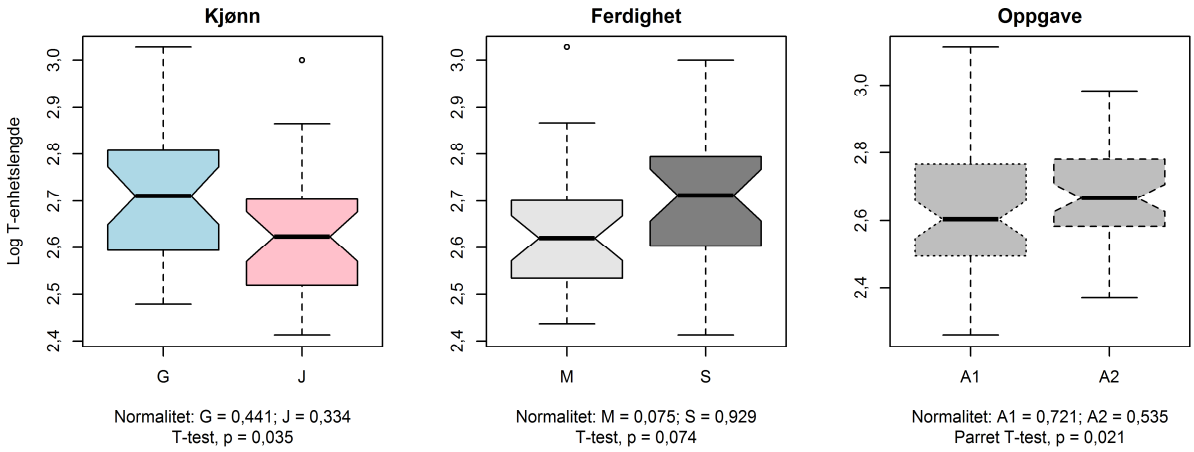
Figur 11-2 viser at fordelingen av t-enhetslengde ikke er utpreget normal, men ganske høyreskjev. Shapiro-Wilks normalitetstest gir $W \approx 0{,}977$, $p < 0{,}05$ for utvalget som helhet ($N = 2x60$), og tilsvarende p-verdier for flere av de aktuelle segmentene. Logaritmetransformasjon øker W-verdiene fra Shapiro-Wilk-testen for alle segmenter unntatt sterke elever, som får høye verdier ($p > 0{,}7$) enten variabelen er transformert eller ikke. Det virker derfor som en rimelig antagelse at t-enhetslengde av natur er mer lognormalfordelt enn normalfordelt, og jeg fortsetter med å regne med logaritmetransformerte verdier av t-enhetslengde.

Det er interessant å vite i hvilken grad elevene har et skrivemønster for t-enhetslengde. De logtransformerte verdiene for henholdsvis håndtekster og tastetekster korrelerer middels sterkt, $R \approx 0{,}41$.



Figur 11-3: Korrelasjon mellom logtransformerte t-enhetslengde i håndtekster og tastetekster. Legg merke til eleven som har påfallende lav verdi i tasteteksten (J5, 205). Uten denne eleven er R ≈ 0,49 (N = 59). Se teksten.

Figur 11-3 viser tendensen tydelig, men figuren viser også at det er en åpenbar utligger, elev 205, som er ei sterk jente. Figuren viser at hun har ganske lange t-enheter i håndteksten (gjennomsnittlig 17,04), men den desidert laveste tastetekstverdien (10,00). Et åpenbart forsøk på løsning av dette potensielle problemet for en parametrisk analyse er å gjøre analysen uten elev 205. Dette resulterer i R ≈ 0,49. `Gvlma` (se 7.2.2.4) rapporterer om ingen problemer med testpremissene enten elev 205 er med i analysen eller ikke. Det virker derfor ganske trygt å anta at det er en middels sterk korrelasjon mellom t-enhetslengde i håndtekstene og tastetekstene, og at elevene har et middels sterkt skrivemønster når det gjelder denne variabelen.



Figur 11-4: T-enhetslengde mot kjønn, ferdighet og oppgave

Figur 11-4 viser at parametrene kjønn og oppgave påvirker t-enhetslengden i utvalget, slik at guttenes tekster har noe lengre t-enheter enn jentetekstene ($d \approx 0{,}54$) og "Ungdomsfylla"-tekstene har noe lengre t-enheter enn "Bøker eller data"-tekstene ($d \approx 0{,}28$). Det er også en svak tendens i utvalget til at de sterke elevene skriver lengre t-enheter enn de middels sterke elevene ($d \approx 0{,}39$).

11.1.1.4 Variansanalyse

I variansanalysen er responsvariabelen differansen mellom logaritmetransformerte t-enhetslengder, mens prediktorene er de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell. Analysen tar utgangspunkt i den maksimale modellen med interaksjoner begrenset til 2 nivåer:

(130) `lm(synD$lTEL~(kjønn+ferdighet+lengde+forskjell)^2)`

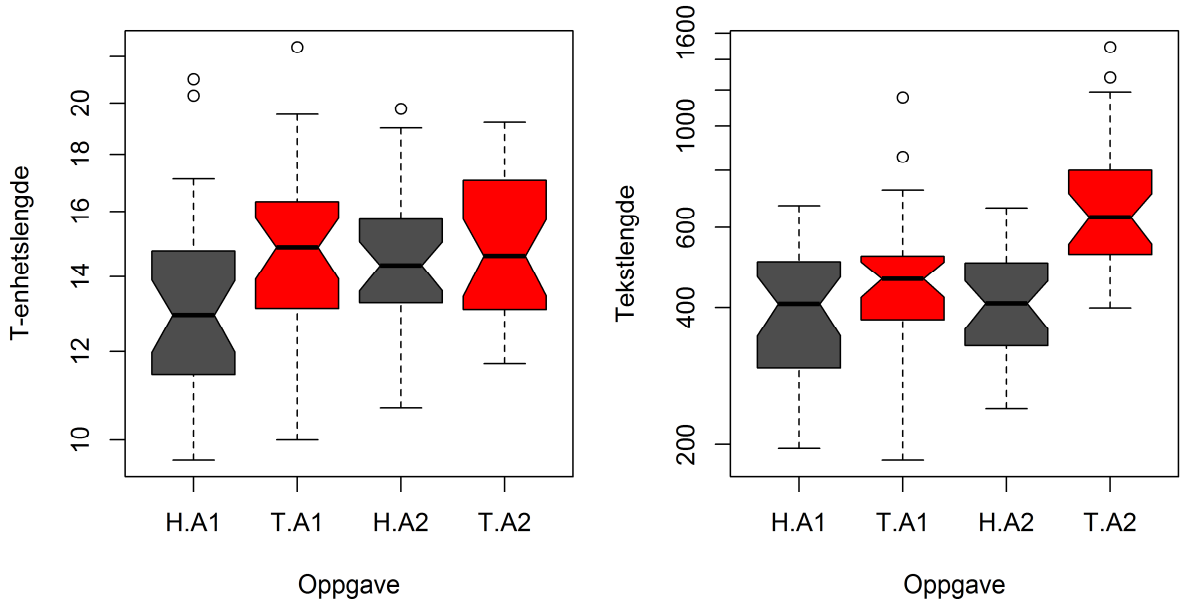
Reduksjon av modellen resulterer i null-modellen som den minimale adekvate modellen.

(131) `lm(formula = synD$lTEL ~ 1)`

```
(Intercept)  0.06533    0.02369   2.758  0.00773 **
---
Residual standard error: 0.1835 on 59 degrees of freedom
```

Denne modellen har $t \approx 2{,}76$, $df = 59$, $p \approx 0{,}0077$, $d \approx 0{,}39$, som er identisk med resultatet fra en paret t-test. (Shapiro-Wilks normalitetstest rapporterer normalitet for differansen mellom tastetekster og håndtekster, $W \approx 0{,}982$, $p \approx 0{,}51$.) Det er altså sterkt signifikant forskjell på t-enhetslengder i håndtekster og tastetekster; t-enhetene er lengre i tastetekstene. Analysen avdekker ikke noen interaksjoner mellom verktøy og de andre parametrene.

Figur 11-5 antyder at oppgaveteksten interagerer med verktøy på en slik måte at verktøyet utgjør en forskjell kun for oppgaven om "Bøker eller data". Det er imidlertid viktig å huske på at den parede designen bryter sammen når oppgavetekst tas med i betraktning. De to boksene for A1-tekstene i figuren representerer ikke de samme elevene.



Figur 11-5: T-enhetslengde: Interaksjon mellom oppgavetekst og skriveverktøy. Sammenlignet med tekstlengde. A1 = "Bøker eller data"; A2 = "Ungdomsfylla". Logaritmiske y-akser.

11.1.1.5 Oppsummering og diskusjon

Tastaturskriving ser ut til å føre til lengre t-enheter, mens det ikke er noen effekter av noen av de fire parametrene i analysen.

Det er vanskelig å tolke denne tendensen uten å gå inn i mer spesifikke analyser av hva som ligger bak lengdeøkningen. Som nevnt kan t-enheter prinsipielt gjøres lengre både ulike måter, nemlig ved å øke antall ledd og ved å øke leddenes lengde. Økningen av leddenes lengde kan skje med subklaususer eller med lengre fraser uten subklaususer. Det er disse ulike prinsippene som utforskes i de følgende delene av dette kapitlet.

Et annet perspektiv er samspillet mellom leksikalske og syntaktiske variabler, og at graden av leksikalsk tetthet og de leksikalske ordenes spesifisitet kan påvirke t-enhetenes lengde. Dette er uansett alltid knyttet til enten leddenes lengde eller antall ledd i t-enheten. Pearsons korrelasjonstester viser at t-enhetslengde korrelerer svært sterkt både med antall leksikalske ord per t-enhet ($R \approx 0.90$) og antall funksjonsord per t-enhet ($R \approx 0.96$). Forskjellen i korrelasjonskoeffisienter synes ubetydelig, noe som støttes av den svake, negative korrelasjonen mellom t-enhetslengde (logaritmetransformert) og leksikalsk tetthet, $R \approx -0{,}16$.

Jeg har også gjort en tilsvarende analyse for gjennomsnittlig antall leksikalske ord per t-enhet. Variansanalysen resulterer da i en nullmodell uten signifikant resultat ($t \approx 1{,}90$, $p \approx 0{,}063$), mens forskjellen mellom middels og sterke elever kommer klarere fram, $d \approx 0{,}64$. Det vil altså si at sterke elever ser ut til å putte mer informasjon inn i hver t-enhet enn middels sterke elever.

11.1.2 Klaususenes lengde

Én av faktorene som kan påvirke t-enhetenes lengde, er lengden av de enkelte klaususene som inngår i t-enheten.

11.1.2.1 Hypotese

Klaususlengde kan økes gjennom å øke antall ledd eller gjennom å gjøre hvert enkelt ledd lengre. Imidlertid er det noe paradoksalt ved målet. Dersom man øker leddlengden på den mest effektive måten, nemlig ved å realisere et ledd som en subklausus, øker ikke dette målet for klaususlengde slik det er definert og regnet ut i denne avhandlingen.

Jeg kjenner ingen sammenlignbare studier som konkluderer om klaususlengde, men Hallidays hypoteser skulle tilsi at klaususlengde er større i skriftlig språk. I så fall er det en naturlig hypotese at klaususer er lengre i håndtekster enn i tastetekster. Chafe og Danielewicz (1987, s. 95-96) finner at lengden av *intonation units* er større i skriftlig språk enn i muntlig, der forfatterne definerer den skriftlige intonasjonsenheten som det samme som en *punctuation unit*. Et segment avgrenset ved tegnsetting er imidlertid ikke det samme som en klausus, og det er uklart i hvilken grad Chafes resultater kan overføres til en analyse om klaususlengde i to varianter av skriftlig språk. På grunnlag av dette er det vanskelig å formulere noen klar hypotese, men man kan kanskje forvente lavere klaususlengde i tastetekstene, på grunnlag av høyere produksjonshastighet. På den andre siden vil redigering kunne øke lengden; en drøfting av faktorenes ulike mekanismer vil være en parallell til tilsvarende drøfting for subklaususfrekvens (se 11.3.1 nedenfor).

Hunt (1965, s. 15) finner stigende klaususlengde med stigende alder for elevene, noe som ikke er særlig oppsiktsvekkende. Når det gjelder de nominelle verdiene som Hunt presenterer, er de ikke direkte sammenlignbare med mine, ettersom de gjelder et annet språk.

11.1.2.2 Korpussøk og utregning

Klaususlengde kan regnes ut som antall ord delt på antall klaususer. Dette innebærer at ord i subklaususer som er ledd i en overklausus, bare regnes som tilhørende én av klaususene. Hvis vi tenker oss at hvert ord alltid regnes som del av klaususen på det laveste nivået, resulterer dette i at klaususer med ledd som fylles av subklaususer, får et kunstig lavt antall ord. For eksempel vil den øverste klaususen i (132), som består av 3 ledd, bare få 2 ord, selv om objektet er obligatorisk.

(132)
```
<t-unit>
Jeg tror
<clause>
  det kan gå utover leseferdighetene våre.
</clause>
</t-unit> [A1-213]
```

Som mål på klaususlengde er ikke nødvendigvis dette et problem, men dersom man skulle teste Hallidays hypotese om at interklausal kompleksitet og intraklausal kompleksitet er motsetninger, vil denne måten å regne klausal lengde på forsterke en negativ korrelasjon mellom de to variablene.

Som generelt mål for klausal lengde velger jeg den enkleste tilnærmingen og operasjonaliserer dette som antall ord dividert på summen av antall klaususer, altså antall subklaususer pluss antall hovedklaususer. Antall hovedklaususer tilsvarer antall t-enheter.

(133) `syn$klL <- lex$n.ordeks / (syn$subkl + syn$TE)`

Siden `syn$TE` også inneholder visse typer fragmenter (4.1.3), resulterer formelen i en noe lavere middelverdi enn reell verdi. På grunn av utregningsmetoden blir altså sekvenser som de understrekede fragmentrestene i (134) – (137), som er de fire første treffene på finitte fragmenter i korpuset, regnet som klaususer:

(134) <t-unit><u>Greit nok</u> at det er bevist at gutter ikke er like glade i å lese bøker som jenter,</t-unit> [A1-222]
(135) <t-unit><u>Viser til leserinnlegget om</u> at gutter bruker data mest, mens jenter heller setter seg ned med en bok.</t-unit> [A1-226]
(136) <t-unit><u>Nei, så hvorfor slenge ut en sånn påstand,</u> som ikke kan stemme i det hele tatt?</t-unit> [A1-233]
(137) <t-unit><u>Jo,</u> fordi data er den nye verden.</t-unit> [A1-237]

Men det finnes også enkelte tilfeller der fragmentresten befinner seg til slutt i sekvensen:

(138) <t-unit>Hvis det er så at de driver med helt andre ting enn det de skal, <u>hva så</u>?</t-unit> [A1-293]

I 5 tilfeller består sekvensen bare av et fragment med finitt verbal uten subklausus, som i (139), og i 2 tilfeller består sekvensen bare av en eller flere (finitte) subklaususer, som i (140). I tilfellene av den siste typen resulterer formelen i en klausus med lengde 0:

(139) <t-unit>Kommer til å bli dritt kjedelig,</t-unit> [A2-233]
(140) <t-unit><clause>Når en gjeng 13-15 år gamle barn stikker ut på fest og drikker seg "fulle" på en halv boks sider og Smirnoff Ice,</clause></t-unit> [A2-210]

Sekvensene varierer altså fra å være svært korte (1 ord) eller helt fraværende (0 ord) til sekvenser av helt normal klaususlengde (8 ord). Dette vil åpenbart påvirke middelverdiene i negativ retning, men i og med at antallet slike fragmenter er såpass lavt i forhold til antall t-enheter, bare 50 av totalt 8544 fullverdige klaususer, blir påvirkningen svært liten, bare i størrelsesorden < 0,05 ord per klausus. For å forenkle prosedyren og gjøre tallene enklere å sammenligne med i senere undersøkelser, har jeg derfor valgt å behandle finitte fragmenter som vanlige t-enheter også i utregningen av gjennomsnittlig klaususlengde.

11.1.2.3 Deskriptiv analyse

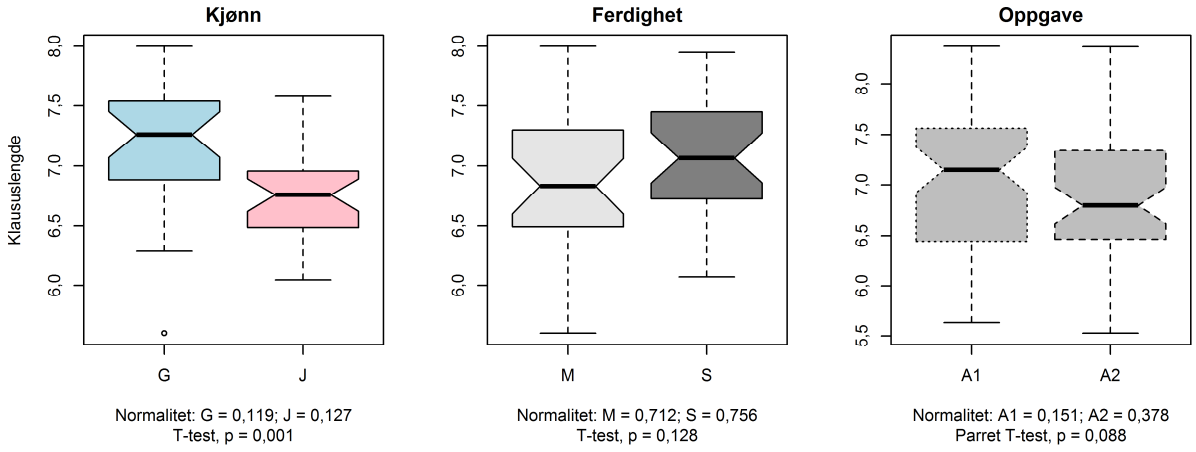
Tabell 11-2 gjengir nøkkeltallene for gjennomsnittlig klaususlengde.

Tabell 11-2: Nøkkeltall for gjennomsnittlig klaususlengde

	middelverdi	median	sd	min	maks
Total	**6,99**	**6,94**	**0,66**	**5,53**	**8,38**
Hånd	6,96	6,94	0,66	5,53	8,37
Tast	7,01	6,94	0,66	5,63	8,38
Middels	6,89	6,80	0,70	5,53	8,37
Sterk	7,09	7,13	0,61	5,63	8,38
Gutt	7,20	7,27	0,68	5,53	8,38
Jente	**6,78**	**6,60**	**0,57**	**5,63**	**8,29**

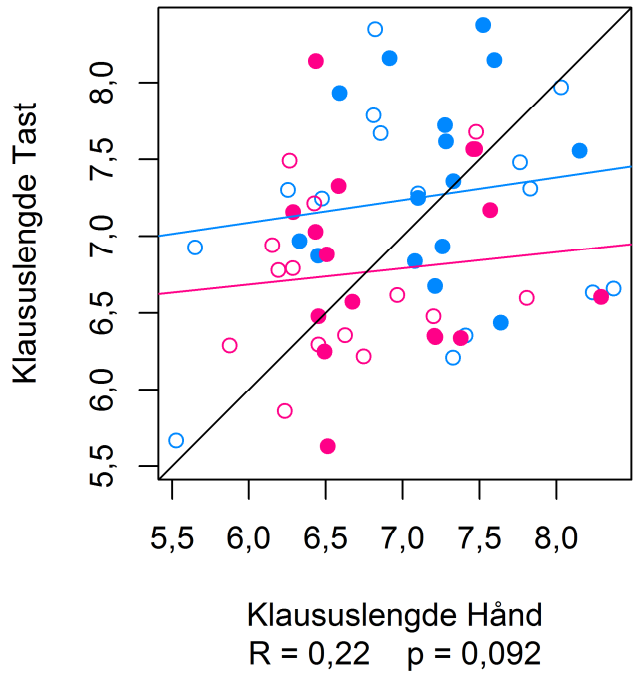
Tabellen viser at gjennomsnittlig klaususlengde ligger rundt 7 ord, med minimums- og maksimumsverdier på henholdsvis ca. 5,5 og 8,4 ord. Standardavviket er i underkant av 0,7. Tabellen indikerer at variasjonen er mindre enn for t-enhetslengde, noe som er naturlig i og med at det finnes færre virkemidler for å variere klaususlengde. Distribusjonen er normal, ifølge Shapiro-Wilks normalitetstest, $W \approx 0{,}979$, $p \approx 0{,}059$. Det er ingen sammenheng mellom gjennomsnittlig klaususlengde og tekstlengde, $R \approx 0{,}08$ med logaritmetransformert tekstlengde.

Figur 11-6 viser at kjønn er den desidert viktigste påvirkningsfaktoren på klaususlengde. Gutter skriver lengre klaususer enn jenter, $d \approx 0.90$. Det er en temmelig markert forskjell. Hverken ferdighet eller oppgave har særlig sterk innvirkning på klaususlengde, noe som kanskje er litt overraskende, men diagrammet viser at de sterke elevene i utvalget produserer noe lengre klaususer enn de middels elevene, som ventet.



Figur 11-6: Klaususlengde mot kjønn, ferdighet og oppgave

Figur 11-7 viser ingen særlig sammenheng mellom gjennomsnittlig klaususlengde i håndtekster og tastetekster, $R \approx 0{,}22$. Det er altså slik at elevene i temmelig liten grad har et etablert skrivemønster når det gjelder klaususlengde.



Figur 11-7: Sammenheng mellom klaususlengde i tastetekster og håndtekster.

11.1.2.4 Variansanalyse

I variansanalysen er responsvariabelen gjennomsnittet av differansen mellom klaususlengder, mens prediktorene er de fire dikotome parametrene kjønn, ferdighet, tekstlengde og tekstlengdeforskjell. Analysen tar utgangspunkt i den maksimale modellen i (141) med interaksjoner begrenset til 2 nivåer:

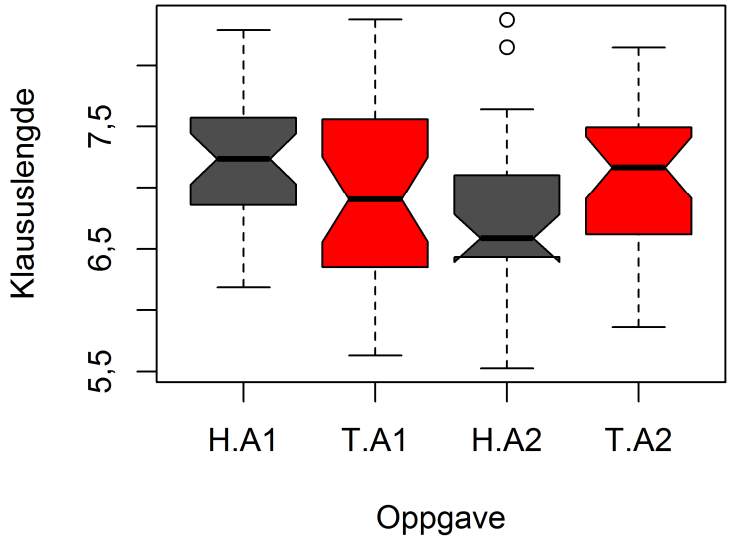
(141)
```
lm (synD$klL~(kjønn+ferdighet+lengde+forskjell)^2)
```

Trinnvis modellreduksjon resulterer i en ikke-signifikant nullmodell, $t \approx 0{,}45$, $p \approx 0{,}65$. (Se appendiks A1.)

(142)
```
lm(formula = synD$klL ~ 1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04796    0.10673   0.449    0.655
---
Residual standard error: 0.8267 on 59 degrees of freedom
```

Verktøy har altså ingen innvirkning på klaususlengde. Det er verdt å merke seg at den markerte kjønnsforskjellen i denne variabelen altså gjelder både hånd- og tastetekster.



Figur 11-8: Klaususlengde. Interaksjon mellom verktøy og oppgave.

Figur 11-8 viser at det er noe interaksjon mellom verktøy og oppgave, og at "Ungdomsfylla"-tekstene særlig er preget av lav klaususlengde i håndtekstene.

11.1.2.5 Oppsummering og diskusjon

Dette avsnittet har vist at kjønn er den eneste parameteren som har innvirkning på gjennomsnittlig klaususlengde. Forskjellen er svært klar og ganske stor, ca. 0,4 ord per klausus, noe som tilsvarer nesten et helt standardavvik ($d \approx 0{,}90$).

Når det gjelder verktøyets innvirkning på klaususlengde, er det ikke avdekket noen effekter. I lys av Chafe og Danielewicz' resultater og Hallidays hypoteser, er dette kanskje mildt overraskende, men det kan være at effektene av produksjonshastighet og redigeringsmuligheter motvirker hverandre fullstendig for denne variabelen. Det kan kanskje også forklares med den svake sammenhengen mellom tastetekster og håndtekster. Selv om variansen i variabelen ikke virker stor, tyder den manglende korrelasjonen på at elevene for klaususlengde ikke har noe klart skrivemønster; dermed blir den tilfeldige variasjonen fra tekst til tekst stor og en målt effekt av situasjonsparametre tilsvarende liten.

Jeg har også gjort tilsvarende analyse på gjennomsnittlig antall leksikalske ord per klausus. Resultatet av variansanalysen blir det samme, mens effekten av oppgave blir klarere, $d \approx 0{,}50$. Se ellers omtalen av antall leksikalske ord per klausus i 10.3.5.

Til slutt vil jeg påpeke at klaususlengde er en variabel der man kanskje skal være forsiktig med å gjøre analyser på gjennomsnittlige verdier. Vi vet at både hovedklaususer og subklaususer kan være svært korte. (132) på side 254 viser et eksempel på en hovedklausus med bare to ord, mens (143) viser en subklausus av samme lengde. I (144) ser vi derimot en klausus med 23 ord, mens t-enheten som helhet har gjennomsnittlig klaususlengde på 10. Dette illustrerer at mye interessant variasjon kan skjule seg bak en teksts gjennomsnittsverdi, og kanskje er frekvensprofiler mer interessant enn gjennomsnittsverdier når det gjelder

lengden på klaususer. En enklere operasjonaliserbart alternativ til frekvensprofiler er antall eller andel svært korte eller svært lange klaususer eller subklaususer.

(143) {Men

{siden det er flere gutter
{som spiller,}}
så har det ingenting å si
{om de leser bøker eller ikke.}} [A1-208]

(144) { {Når jeg gikk i 10.-klasse,}

opplevde jeg
{at mange i min klasse valgte å lese referat fra boken på internett
i stedet for å sette seg ned og lese bøker.}} [A1-212]

11.1.3 Korte subklaususer

Som jeg peker på i 11.1.2.5, er gjennomsnittsverdier av klaususlengde eller subklaususlengde en variabel som potensielt kan skjule mye interessant variasjon. Alternative variabler kan være å måle variasjonen i subklaususlengde, for eksempel med entropi, eller tettheten eller frekvensen av svært korte eller svært lange subklaususer. Hvis man skal måle antall lange subklaususer, må man også vurdere nærmere hvordan subordinering innenfor subklaususer skal behandles, og jeg har valgt en løsning som er enklere operasjonaliserbar, nemlig å måle antall subklaususer som består av 1, 2 eller 3 ord uten intern subordinering. I dette avsnittet lar jeg dette være definisjonen på det jeg kaller "korte subklaususer".

11.1.3.1 Hypotese

Jeg kjenner ingen undersøkelser som spesifikt har brukt denne variabelen, men det virker naturlig basert både på Hallidays hypoteser og på forutsetningen om at i hvert fall en del typer utbrytninger er mer frekvente i muntlig språkbruk (Fjeld, 2007, s. 54-55), å hypotetisere at korte subklaususer er mer frekvente i tastetekstene.

11.1.3.2 Korpussøk

Korte subklaususer uten underordning gjenfinnes med søket i (145).

(145)
```
<clause>{([lemma="\$.*"]|[<corr>])*([!<clause> &
!</clause>]([lemma="\$.*"]|[</corr>])*){1,3}</clause> [414 treff]
```

Ettersom søket etterspør et spekter av lengder fra 1 ord og oppover, er det ikke nødvendig å spesifisere at de etterspurte ordene ikke skal være skilletegn. Søket er "grådig" og vil finne det maksimale antall treff, noe som innebærer å finne alle subklaususer som faktisk inneholder inntil 3 ord.

11.1.3.3 Utregning

Om man skal beregne frekvens av slike korte subklaususer, er spørsmålet om målestokk aktuelt. Man kan bruke samme målestokk som for subklaususfrekvens (se 11.3.1 nedenfor), altså gjennomsnittlig antall korte subklaususer per t-enhet. Men man kan også måle andelen av subklaususer som er korte, altså bruke antall subklaususer som målestokk. Om man bruker andel av subklaususer, får ikke nødvendigvis tekster med mange korte subklaususer høye verdier, dersom tekstene også har høy frekvens av subklaususer generelt. På den annen side kan man også tenke seg at tekster med mange korte subklaususer også har mange korte t-enheter uten subordinering, som dermed bidrar til lavere verdier for antall subklaususer. En mer teorinøytral målestokk, som tekstlengde i antall ord, kan i dette tilfellet kanskje reflektere bedre det inntrykket leseren får av teksten, i form av hvor "ofte" man som leser møter en kort subklausus.

Spørsmålet om målestokk er ikke enkelt å besvare, og for denne variabelen har jeg forsøkt å gjennomføre analysen med alle de tre nevnte målestokkene, altså antall ord, antall subklaususer og antall t-enheter. Resultatet er metodisk interessant.

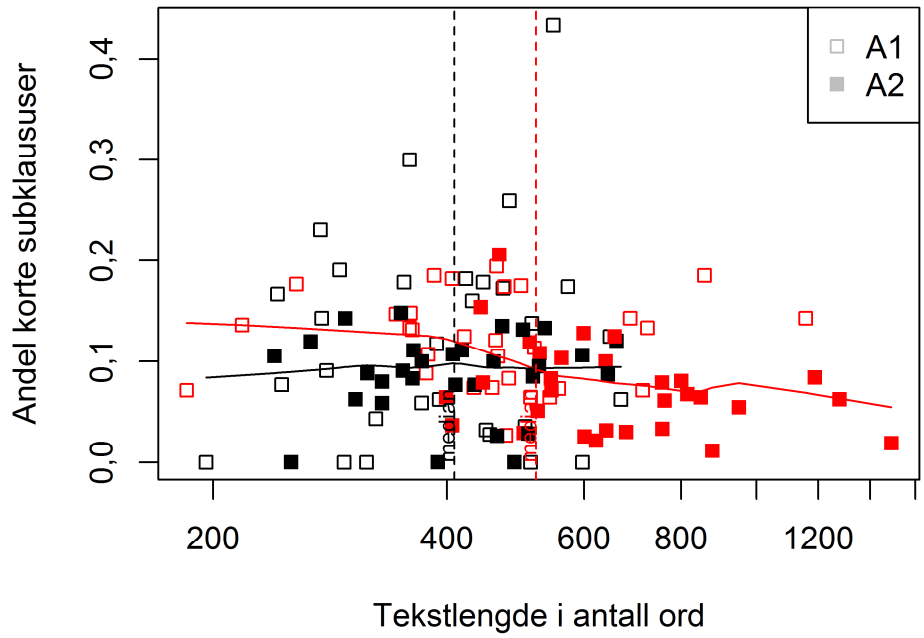
11.1.3.4 Deskriptiv analyse

Jeg presenterer her nøkkelverdiene for bare én av variantene, nemlig andel subklaususer som er korte (tabell 11-3). Totalt 8 tekster har ingen korte subklaususer, og som det går fram av tabellen, er alle disse håndtekster. Gulveffekten gir en viss høyreskjevhet i variabelen, $W \approx 0{,}921$, $p < 0{,}001$, ifølge Shapiro-Wilks normalitetstest.

Tabell 11-3: Nøkkeltall for andel korte subklaususer

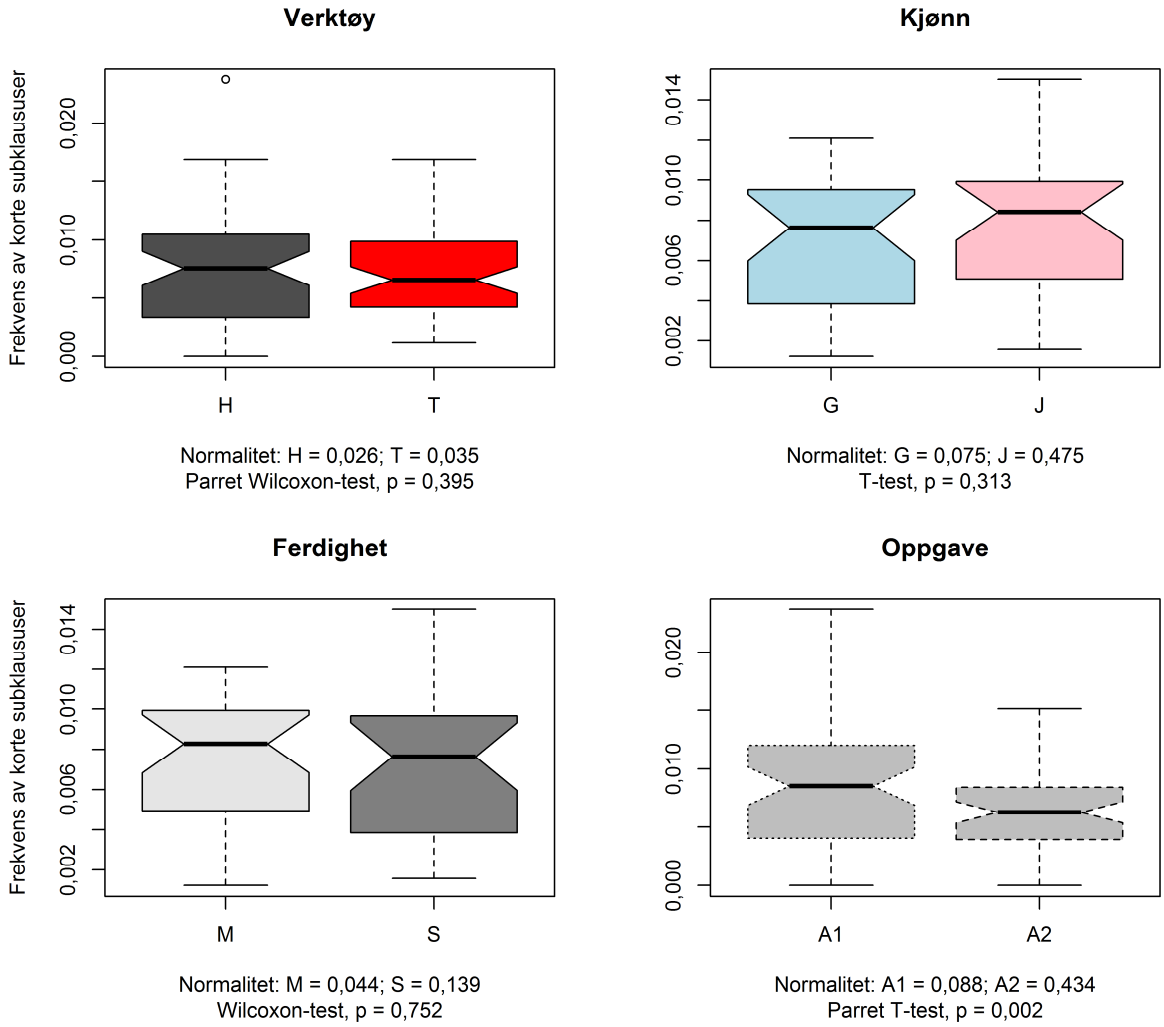
	middelverdi	median	sd	min	maks
Total	**0,101**	**0,091**	**0,066**	**0**	**0,433**
Hånd	0,104	0,098	0,079	0	0,433
Tast	0,097	0,084	0,051	0,012	0,206
Middels	0,101	0,098	0,054	0	0,206
Sterk	0,101	0,084	0,077	0	0,433
Gutt	0,096	0,093	0,062	0	0,300
Jente	**0,106**	**0,090**	**0,071**	**0**	**0,433**

Figur 11-9 viser også at det ser ut til å være en negativ korrelasjon mellom tekstlengde og andel korte subklaususer i tastetekstene; Pearsons korrelasjonstest viser en middels sterk korrelasjon, $\text{rho} \approx -0{,}34$.



Figur 11-9: Andel korte subklaususer og sammenheng med tekstlengde

Alle tre variantene av variabelen viser klart lavere verdier for "Ungdomsfylla"-tekstene, men ingen vesentlige forskjeller for verktøy, kjønn eller ferdighet. Figur 11-10 viser effektene for frekvens av korte subklaususer per antall ord.



Figur 11-10: Frekvens av korte subklaususer per ord, fordelt etter skriveverktøy, kjønn, skriveferdighet og oppgave

11.1.3.5 Variansanalyse

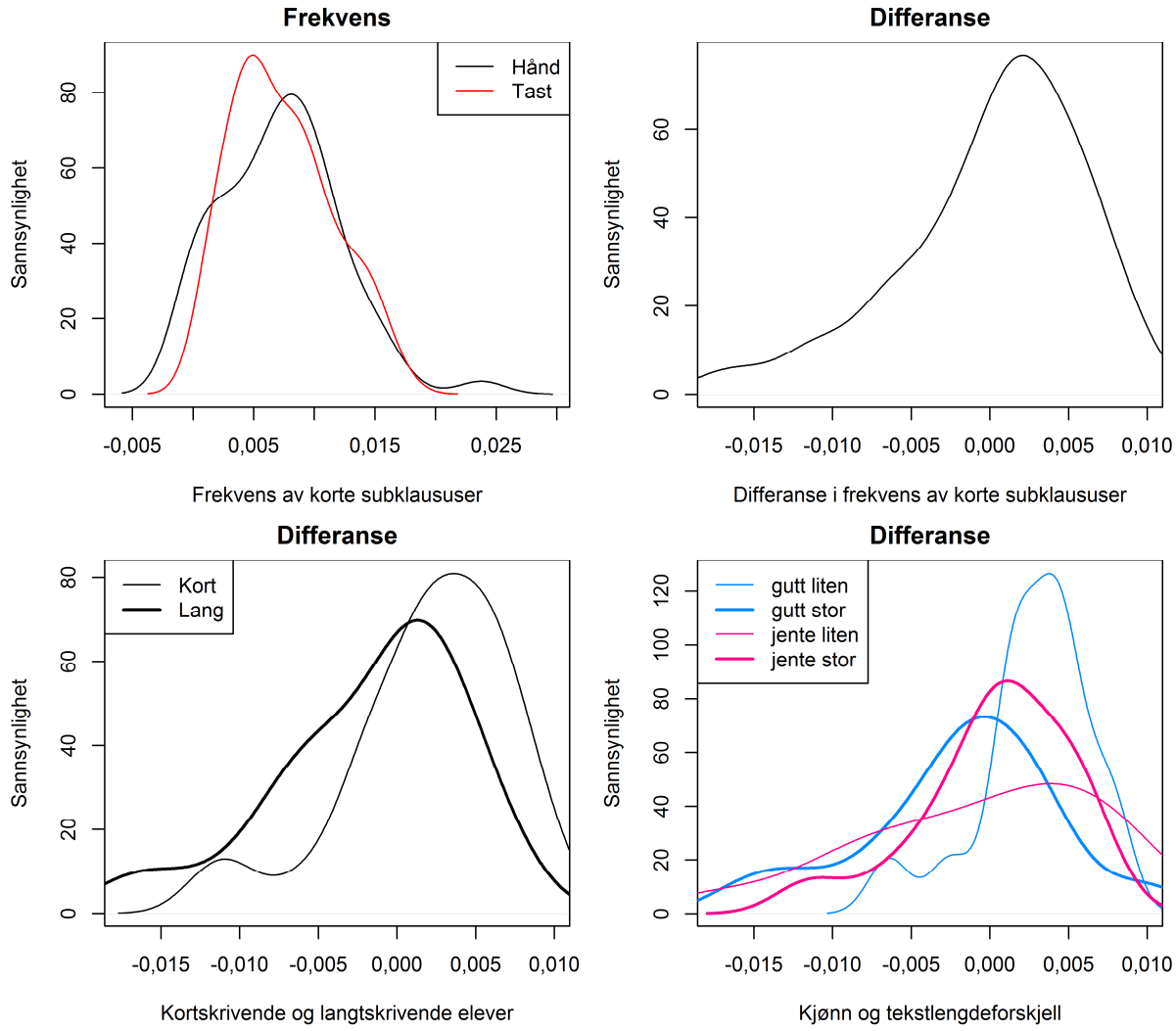
Jeg har gjennomført variansanalyse for alle tre variantene av variabelen, med utgangspunkt i de maksimale modeller som er gjengitt nedenfor sammen med de respektive anova-tabellene. (146) viser frekvens per antall ord, (147) viser tetthet per antall subklaususer, og (148) viser frekvens per antall t-enheter. For tetthet per antall subklaususer er responsvariabelen differansen av logit-transformerte verdier; for de to andre er det differansen av frekvenstall som blir brukt som responsvariabel. Det interessante er at analysen gir ganske ulike resultater for de tre variantene.

(146) `lm(synD$subklL_3F~(kjønn+ferdighet+lengde+forskjell)^2)`

```
                 Df    Sum Sq   Mean Sq F value Pr(>F)
kjønn             1 0.0000065 6.460e-06   0.210 0.6488
lengde            1 0.0002028 2.028e-04   6.589 0.0130 *
forskjell         1 0.0000172 1.718e-05   0.558 0.4583
kjønn:forskjell   1 0.0001409 1.409e-04   4.577 0.0369 *
```

```
Residuals       55 0.0016932 3.078e-05
---
Multiple R-squared:  0.1783,    Adjusted R-squared:  0.1185
F-statistic: 2.983 on 4 and 55 DF,  p-value: 0.02666
```

For analysen i (146) av frekvens per antall ord i teksten rapporterer `gvlma` (se 7.2.2.4) at premissene for anova-analysen ikke er oppfylt; for skjevhet er p ≈ 0,039. (Se appendiks A4.) Selv om t-test og anova er potensielt ømfintlig for skjevhet, er risikoen for feil liten når utvalgene ikke er for små, utvalgene er like store, skjevheten er moderat og likeformet i utvalgene (7.2.3.1). Differansene i variabelen er moderat venstreskjev, og de aktuelle utvalgene er like store, enten 30 eller 15. Figur 11-11 viser at skjevheten er moderat og omtrent likeformet i de ulike utvalgene.



Figur 11-11: Tetthetskurver som viser venstreskjevhet i differansen av korte subklaususer per antall ord

Jeg mener på bakgrunn av dette at resultatene fra anova-analysen av korte subklaususer per antall ord kan aksepteres som gyldig.

(147) `lm(synD$subklL_3.subkl~(kjønn+ferdighet+lengde+forskjell)^2)`

```
                          Df Sum Sq Mean Sq F value Pr(>F)
```

```
ferdighet            1   1.75   1.745   1.633 0.2066
forskjell            1   3.68   3.679   3.442 0.0688 .
ferdighet:forskjell  1   4.45   4.448   4.162 0.0461 *
Residuals           56  59.86   1.069
---
Multiple R-squared:  0.1416,    Adjusted R-squared:  0.0956
F-statistic: 3.079 on 3 and 56 DF,  p-value: 0.03472
```

`Gvlma` (se 7.2.2.4) viser at premissene for anova-analysen er oppfylt (se appendiks A4).

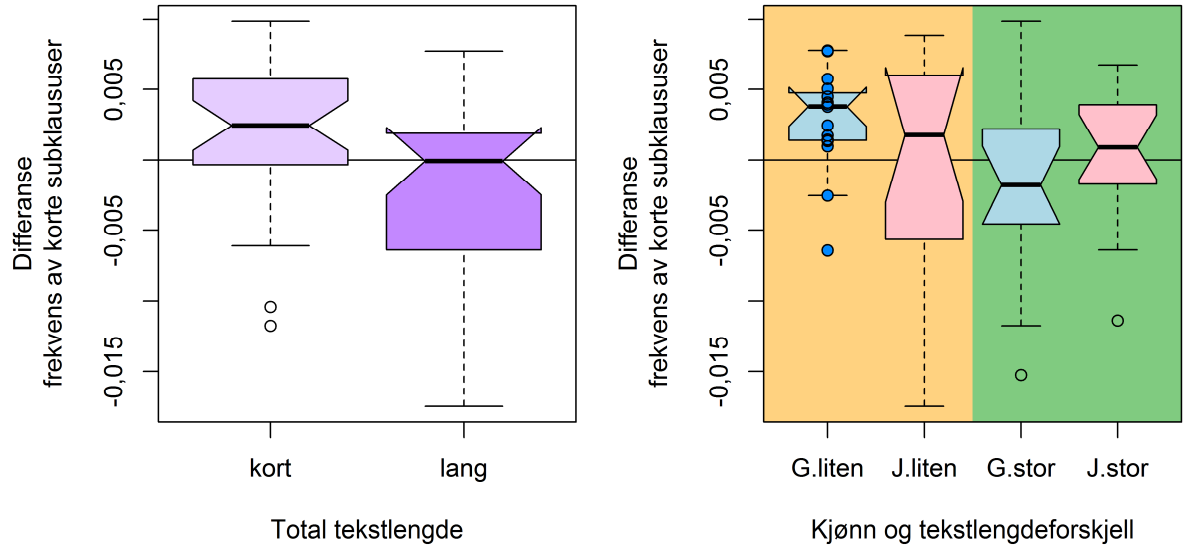
(148) `lm(synD$subklL_3.te~(kjønn+ferdighet+lengde+forskjell)^2)`

```
            Df Sum Sq  Mean Sq F value Pr(>F)
ferdighet    1 0.0282 0.028162   4.797 0.0325 *
Residuals   58 0.3405 0.005871
---
Multiple R-squared:  0.07639,   Adjusted R-squared:  0.06046
F-statistic: 4.797 on 1 and 58 DF,  p-value: 0.03255
```

`Gvlma` (se 7.2.2.4) viser at premissene for anova-analysen er oppfylt (se appendiks A4.)

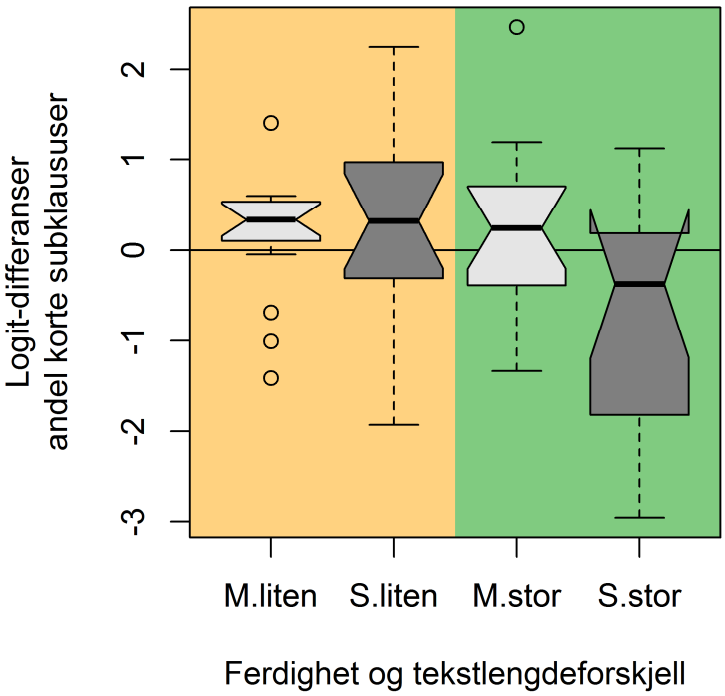
De tre analysene gir altså alle signifikante resultater, men de frambringer ulike prediktorer som signifikante.

Resultatene for antall korte subklaususer per antall ord er den mest komplekse. Variansanalysen viser svak signifikans for prediktoren total tekstlengde ($p < 0{,}05$, $d \approx 0{,}67$), men dessuten svak signifikans for interaksjonen mellom kjønn og forskjell i tekstlengde. Elever som generelt skriver korte tekster, bruker flere korte subklaususer i tastetekstene, mens den mest iøynefallende tendensen i interaksjonen er at gutter som har liten forskjell i tekstlengde, bruker flere korte subklaususer i tastetekstene, mens gutter som skriver mye lengre på tastatur, bruker færre korte subklaususer i tastetekstene. Som figur 11-12 viser, er det bare to gutter i den første kategorien som bruker færre korte subklaususer i tastetekstene. Imidlertid oppgir Tukeys HSD-test ingen av enkeltkontrastene i interaksjonen som signifikante, selv om interaksjonen som helhet altså er det (se appendiks A5). Forskjellen mellom de to guttesegmentene blir oppgitt til $p \approx 0{,}19$ ($d \approx 0{,}93$). Det er imidlertid *interaksjonen* mellom kjønn og tekstlengdeforskjell som anova rapporterer som signifikant, og boksdiagrammet viser tydelig hvordan forskjellen mellom gutter og jenter med liten tekstlengdeforskjell (til venstre i diagrammet) er forskjellig fra forskjellen mellom gutter og jenter med stor tekstlengdeforskjell (til høyre i diagrammet). Denne typen interaksjoner er Tukeys HSD-test ikke så godt egnet til å avdekke, og jeg hevder at det er en forskjell mellom kjønnene her, selv om Tukeys post-hoc-test ikke rapporterer noen forskjell mellom noen av de fire segmentene.



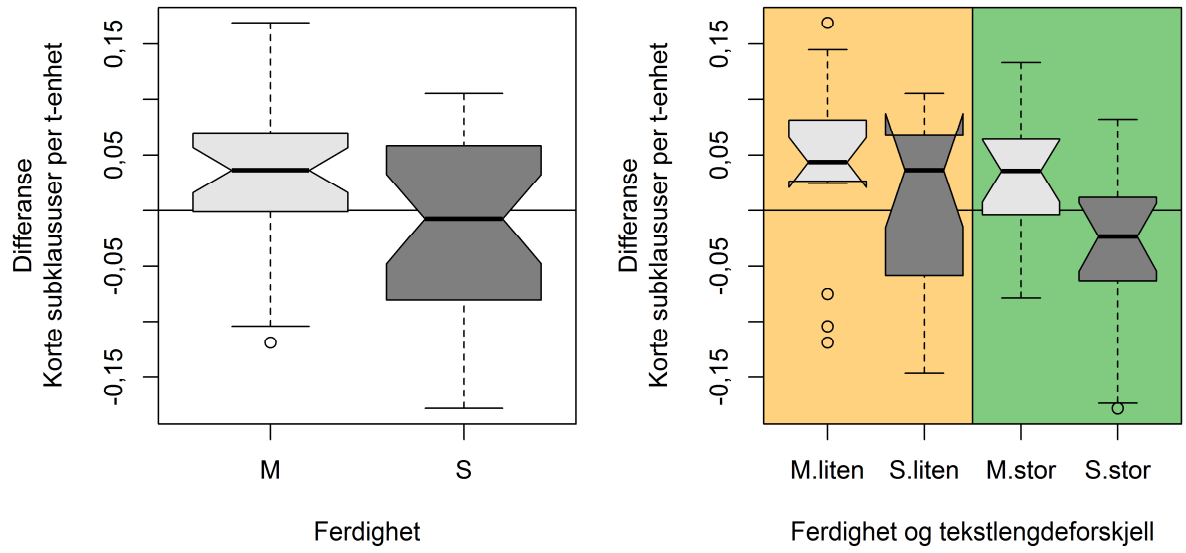
Figur 11-12: Boksdiagrammer som viser resultatene av anova-analysen for frekvens av korte subklaususer per antall ord

Resultatet for tetthet av korte subklaususer, altså andelen av subklaususene som er korte, gir et helt annet mønster. Denne analysen viser svak signifikans for interaksjonen mellom ferdighet og tekstlengdeforskjell, mens hverken total tekstlengde eller kjønn er relevante. Tukeys HSD-test viser at forskjellen mellom sterke elever som skriver mye lengre på tastatur, og sterke elever som ikke gjør det, er svakt signifikant ($p < 0{,}05$, $d \approx 0{,}89$, se appendiks A5), og figur 11-13 viser at de fleste sterke elever med stor tekstlengdeforskjell har en tendens til å bruke færre korte subklaususer i tastetekstene, mens resten av elevgruppen har den motsatte tendensen.



Figur 11-13: Resultatet av anova-analysen for tetthet av korte subklaususer per antall subklaususer

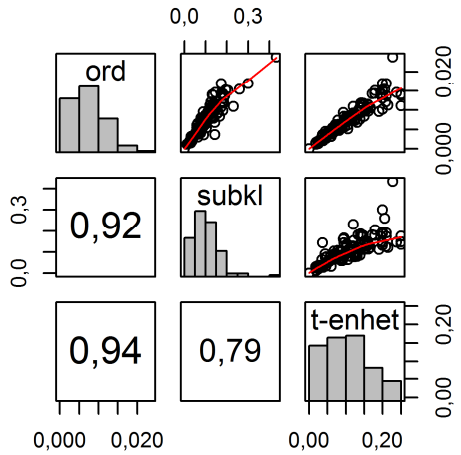
Resultatet for frekvens av korte subklaususer per t-enhet ligner på det foregående resultatet. Denne analysen viser svak signifikans for skriveferdighet alene, og ingen interaksjoner. Diagrammet til venstre i figur 11-14 viser at middels elever har en tendens til å bruke flere korte subklaususer i tastetekstene, mens det ikke er noen slik tendens for de sterke elevene. Dette er altså ikke det samme resultatet som i den forrige analysen, men vi ser at de to analysene har klare likhetstrekk i hva som er forskjellen mellom middels og sterke elever. Diagrammet til høyre viser at interaksjonen mellom ferdighet og tekstlengdeforskjell viser de samme tendensene som for andel korte subklaususer, selv om interaksjonen her ikke er signifikant.



Figur 11-14: Boksdiagram som viser resultatet fra anova-analysen for frekvens av korte subklaususer per t-enhet (til venstre), og den ikke-signifikante interaksjonen mellom ferdighet og tekstlengdeforskjell (til høyre).

Det er også likhetstrekk mellom den siste analysen og ett av delresultatene for frekvens av korte subklaususer, ettersom det er noe korrelasjon mellom ferdighet og total tekstlengde; 19 middels elever skriver kort, mens 11 skriver langt.

Disse tre resultatene illustrerer den innvirkning valg av målestokk kan ha på resultatene, og minner oss også om å fortolke andre resultater med en viss varsomhet. Som figur 11-15 viser, er det sterke korrelasjoner mellom de tre variabelvariantene, noe svakere mellom ett av parene enn mellom de to andre. Likevel blir altså resultatene fra de tre analysene ganske forskjellige og åpner for ulike fortolkninger.



Figur 11-15: Korrelasjoner mellom de tre variantene av variabelen for korte subklaususer som er beskrevet i teksten

I visse tilfeller kan man hevde at bruk av forskjellige målestokker kan frembringe forskjellige variabler. Se for eksempel diskusjonen om leksikalsk tetthet i 9.2.1. Men om man analyserer samme variabel på tre forskjellige måter, slik jeg har gjort her, kan man ikke tillegge p-verdiene samme tyngde som ellers, men må justere dem for repetert testing (FWER). Man kan altså ikke bruke disse analysene til å hevde at man har funnet signifikante resultater for både tekstlengde, interaksjon mellom kjønn og tekstlengdeforskjell, interaksjon mellom skriveferdighet og tekstlengdeforskjell, og skriveferdighet.

Formålet mitt med å utføre de tre analysene er å illustrere målestokkens potensielle innvirkning på analyseresultatene, og ikke å frambringe flere analyseresultater. I analysen av korte subklaususer har jeg valgt å fokusere på den varianten som bruker antall ord i teksten som målestokk. Se diskusjonen i 11.1.3.6 under og 11.1.3.3 over.

11.1.3.6 Oppsummering og diskusjon

Analysene av korte subklaususer demonstrerer at spørsmålet om hva det betyr at et trekk er frekvent, ikke er trivielt. Prinsippet om å anvende mulige brukskontekster som målestokk er godt for variabler som har klart avgrenset semantikk og grammatisk bruksområde, men for variabler med mer komplekse egenskaper er det ikke enkelt å avgjøre hva som er "mulige brukskontekster". Dessuten kan det være vanskelig å gjenfinne de relevante kontekstene automatisk i et korpus.

"Mulige brukskontekster" for korte subklaususer er neppe subklaususer generelt. Det er ikke slik at språkbrukeren bestemmer seg for å bruke en subklausus og deretter avgjør om den skal være kort eller lang. Mulige brukskontekster for korte subklaususer er heller ikke t-enheten, ettersom hver t-enhet kan inneholde flere korte subklaususer, og det neppe er mulig å avgrense hvor mange korte subklaususer som kan settes inn i en t-enhet.

Om man skulle forsøke å konstruere en oversikt over mulige kontekster for korte subklaususer, måtte man trolig ta for seg en subklausustype ad gangen. Korte relativklaususer kan i hovedsak opptre sammen med nomenfraser. Imidlertid er dette trolig et for snevert perspektiv, ettersom relativklaususer også brukes for eksempel i forskjellige utbrytningskonstruksjoner for å fokusere eller kontrastere et ledd, som vist i (150) – (153).

(149) Jon har gjort det. [BUJ]
(150) Det er Jon som har gjort det. [BUJ]
(151) Den som hadde gjort det, var Jon. [BUJ]
(152) Det er det Jon som har gjort. [BUJ]
(153) Gjort det er det Jon som har. [BUJ]

Riktig nok kan relativklaususen i alle eksemplene sies å være knyttet til \Jon\, men disse utelukker jo ikke at andre relativklaususer knyttes til nomenfrasen:

(154) Jon, som ellers er en grepa kar, har gjort det. [BUJ]
(155) Det er Jon, som ellers er en grepa kar, som har gjort det. [BUJ]

På lignende måte kan et prinsipielt uavgrenset antall relativklaususer knyttes til hver nomenfrase, og nomenfrase kan ikke regnes som mulig brukskontekst for korte relativklaususer.

(156) Jon, som er en grepa kar, som bor i Moss, og som tidligere ikke har gjort en flue fortred, har gjort det. [BUJ]

Korte nominalklaususer kan drøftes på lignende måte. Nominalfraser kan stå som nominale ledd, for eksempel subjekt, objekt og preposisjonsutfylling. Imidlertid er det bare et lite utvalg av nominale ledd som potensielt kan fylles av en nominalklausus, og de potensielle anvendelseskontekstene er nok heller forankret i den semantiske strukturen enn den syntaktiske. Også nominalklaususer kan anvendes for utbrytning.

(157) Det syntes jeg han skulle gjøre. [BUJ]

Når det gjelder adverbialklaususer, kan disse danne et stort spekter av forskjellige typer adverbialledd, og det er vanskelig å avgrense hva slags og hvor mange adverbiale ledd en t-enhet eller en klausus kan ha.

Generelt er det umulig å bruke syntaktiske eller leksikalske enheter som "mulig brukskontekst" for subklaususer, og man må i praksis bruke en målestokk som er begrunnet på en annen måte, og helst i så stor grad som mulig en "nøytral" størrelse som ikke interagerer med det man forsøker å måle.

- antall ord
- antall ord av en viss type
- antall fraser av en viss type
- antall ledd
- antall ledd av en viss type
- antall klaususer
- antall klaususer av en viss type
- antall t-enheter

Generelt mener jeg at de mer avanserte målestokkene er problematiske fordi de kan skape interaksjonseffekter. Et typisk eksempel er dersom man bruker antall t-enheter som målestokk for en variabel som korte subklaususer, Hvis en tekst har mange korte t-enheter, som kan være knyttet til en helt annen tekstlig egenskap enn den som skal måles, kan det senke det relative målet for variabelen, mens lange, komplekse t-enheter av helt andre typer enn den relevante kan ha motsatt effekt. Tilsvarende problemer kan oppstå om man bruker antall subklaususer som målestokk. Hvis en tekst har noen få korte subklaususer men heller ikke særlig mange lengre subklaususer, vil den få en høy tetthetsverdi selv om antall forekomster av trekket er ganske lavt. Samme type argumenter kan brukes ved målestokker av andre typer, for eksempel en viss type ledd eller en viss type ord.

Jeg tror derfor at det er gode grunner til å opprettholde en mer teorinøytral målestokk som antall løpeord, selv om de analyseapparatene som er tilgjengelige for oss i dag, gir oss andre

muligheter enn for eksempel Biber (1988) hadde. Å bruke samme målestokk på alle eller de fleste variablene medfører også at det virker mer forsvarlig å putte dem inn i en multivariat analyse som prinsipalkomponentanalyse (se kapittel 12.)

Visse variabler synes dog å ha egenskaper som gjør det svært naturlig å måle dem som andeler eller forholdstall, som for eksempel andelen t-enheter med ett ord i forfelt (11.3.2), og jeg har valgt å beholde *andel* som mål for denne variabelen, og dessuten for adverbiale subklaususer (11.2.2). Se også diskusjonene om og preposisjoner (11.2.1) og subklaususfrekvens (11.3.1). Jeg mener likevel at også disse valgene kan diskuteres, og at det kanskje ville være fornuftig å bruke antall løpeord som målestokk også for disse, eventuelt bare for enkelte av dem.

11.2 Antall ledd per klausus

En måte å øke den gjennomsnittlige klaususlengden på er å øke antall ledd per klausus.

Den grammatiske taggingen i korpuset er ikke nøyaktig nok til å kunne måle antall ledd i klaususer med noen særlig nøyaktighet. Det går imidlertid an å nærme seg denne variabelen mer indirekte gjennom å måle syntaktiske egenskaper som man kan gå ut fra har sammenheng med hvor mange ledd som finnes i klaususen.

Antall ledd i en klausus kan være relatert til hovedverbets valens, altså hvor mange semantiske roller verbalet deler ut, og det virker rimelig å hevde at et verbal med høyere valens tilfører klaususen mer kompleksitet. Men i tillegg til de utdelte rollene, kan klaususens kompleksitet også økes gjennom adjungerte ledd.

Jeg har under arbeidet med dette kapitlet forsøkt en tilnærming til ledd generelt, uavhengig av om de er utfyllinger til verbalet eller adjungerte ledd. Dette har jeg gjort gjennom analyser av frekvens av substantiv, pronomen, nominale subklaususer og adverbiale subklaususer. Deretter nærmet jeg meg antall tildelte roller gjennom å se på forekomster av det enkle verbet \VÆRE\ som hovedverb. Ingen av disse analysene gav positive funn, og man kan mistenke at en viktig årsak til dette rett og slett er at de valgte variablene er svake operasjonaliseringer av de egenskaper jeg søkte å undersøke.

Jeg konsentrerer diskusjonen om antall ledd rundt preposisjonsfraser og adverbiale subklaususer som indirekte mål på adjungerte ledd i t-enhetene, og det er disse to variablene dette delkapitlet omhandler.

11.2.1 Preposisjonsfraser

Antall preposisjonsfraser per klausus er et forsøk på en delvis operasjonalisering av antall adjungerte ledd i klaususen. Det virker sannsynlig at antall ikke-obligatoriske ledd er et kompleksitetsmål som er mer direkte påvirkelig av både kognitive begrensninger og redigering enn antall obligatoriske ledd, så potensialet for påvirkning fra skriveverktøyet synes større enn for antall obligatoriske ledd. Selvfølgelig er det også slik at endel

preposisjonsfraser representerer obligatoriske ledd, men det er liten grunn til å tro at dette vil forstyrre eventuelle effekter av skriveverktøyet på antall adjungerte ledd, annet enn ved økt tilfeldig variasjon i materialet. Det er også slik at alle preposisjonsfraser ikke representerer selvstendige ledd, men de øker i alle fall kompleksiteten i klaususen.

11.2.1.1 Hypotese

(Biber, 1988, s. 102) har preposisjoner som en negativ faktor (vekt –0,54) i dimensjon 1, altså et uttrykk for en mer integrert og mindre involvert teksttype. Chafe og Danielewicz (1987, s. 98-99) finner også mer preposisjonsfraser i planlagt, skriftlig språk, selv om de peker på at noen preposisjonsfraser er tettere knyttet til verbet, og at disse er mer frekvente i spontan, muntlig språkbruk. De peker på behovet for mer forskning på ulike typer preposisjonsfraser. Halliday (2004, s. 654-656) viser til at kompleksiteten i skriftlig språk typisk er knyttet til høyere antall leksikalske ord per klausus, mens kompleksiteten i muntlig språk er knyttet til høyere frekvens av klaususer og lavere antall leksikalske ord per klausus. Uten at han nevner preposisjoner eller preposisjonsfraser eksplisitt, er det likevel klart at preposisjonsfraser er en måte å øke *antall* leksikalske ord i klaususen, mens effekten på *andelen* leksikalske ord er mer usikker.

Selv om Chafe og Danielewicz tar viktige forbehold, peker resultatene i forrige avsnitt alle i samme retning, nemlig at preposisjonsfraser først og fremst representerer en type planlagt, integrert kompleksitet som vi venter å finne mest av i håndtekstene. Et annet viktig forbehold i overføringen av hypoteser fra studiene av Biber og av Chafe og Danielewicz, er at begge disse bruker antall løpeord som målestokk, i motsetning til antall klaususer, som er brukt i denne studien. Som demonstrert i 11.1.3.5 er ikke antall preposisjoner per løpeord og antall preposisjoner per klausus nødvendigvis den samme variabelen.

11.2.1.2 Korpussøk og utregning

Preposisjonsfraser er ikke direkte tilgjengelig gjennom korpusfunksjonaliteten, men må gjenfinnes indirekte ved søk etter preposisjoner. Preposisjoner er tagget i korpuset og enkelt tilgjengelig gjennom søket i (158):

(158) `[features=("prep")]`
(159) `<clause>[]?[features=("prep") & word="som"%c]`

I tillegg er det slik at 26 tilfeller av totalt 570 forekomster av \av\ er feiltagget som forkortelsen \AV\ og tagget som `adj` i stedet for `prep`. Disse 26 tilfellene burde ha vært korrigert manuelt, men de er ikke medregnet i analysen. Jeg har gått manuelt gjennom de 29 forekomstene i trefflista for `prep` som har subjunksjon som den prioriterte taggen, og fjernet 18 treff som er subjunksjoner og ikke preposisjoner. Jeg har også fjernet 1 treff på \for\ der \for\ fungerte som konjunksjon.

En del ord som omfattes av søket, er imidlertid ord som vanligvis kategoriseres som adverb, og som ikke danner preposisjonsfraser med utfylling, slik vi ser et eksempel på i (160). Det er dessuten en del tilfeller der subjunksjonen \som\ er tagget som preposisjon, som i

eksemplet i (161). Det er mulig at søket etter preposisjonsfraser kunne vært gjort mer nøyaktig ved å kombinere det med taggen for preposisjonsutfylling, `@<p-utfyll`, men dette er heller ikke trivielt, og jeg har ikke gjort noe forsøk på dette. Disse forekomstene realiserer jo dessuten uansett et ledd, som er det underliggende formålet med søket. I tillegg kunne de fleste tilfeller av \som\ som subjunksjon enkelt vært fjernet fra søket ved å fjerne de tilfeller av \som\ som opptrer ved subklaususstart, som i (159). I eksemplet i (161) representerer subklaususen et eget ledd, men i de fleste tilfellene gjør ikke relativklaususene det. Jeg har ikke fjernet disse 164 subjunksjonene fra trefflista med preposisjoner, og det er en svakhet ved metoden. Men i forhold til de 7606 forekomstene som er med i analysen, utgjør de bare en ganske liten feilkilde.

(160) Her står det at jenter ikke driver med data, men leser bøker. [A1-206]
(161) Som jeg sa tidligere i dette innlegget, så kan jeg ikke huske når jeg sist leste en bok. [A1-202]

Antall preposisjoner per klausus regnes ut med følgende formel:

(162)
```
pos$prep.kl <- pos$prep / syn$kl
```

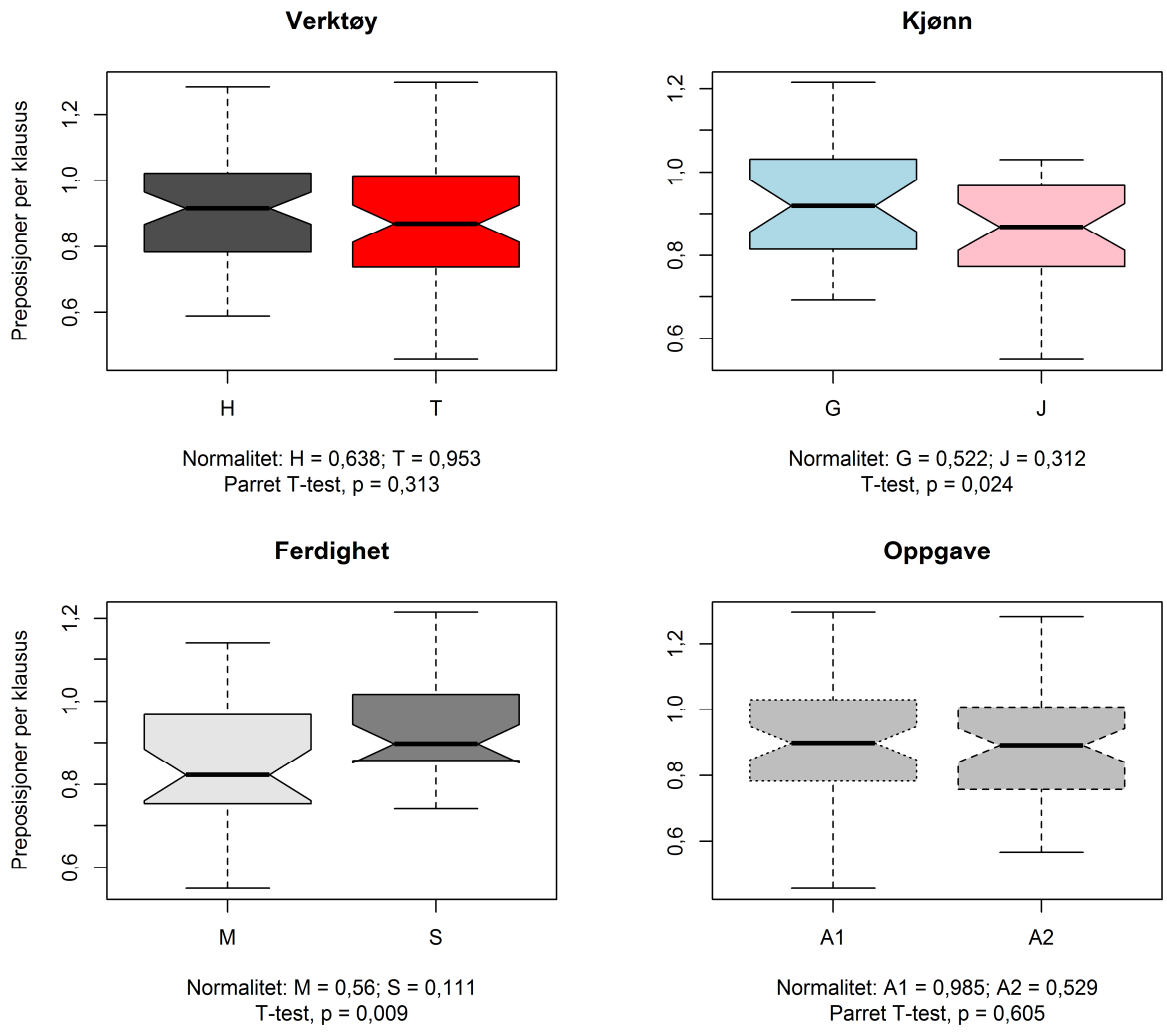
11.2.1.3 Deskriptiv analyse

Tabell 11-4 viser nøkkeltallene for antall preposisjonsfraser per klausus. Tabellen viser at gjennomsnittsverdiene ligger litt under 1, med et standardavvik på 0,17.

Tabell 11-4: Nøkkeltall for preposisjonsfraser per klausus

	middelverdi	median	sd	min	maks
Total	**0,893**	**0,896**	**0,168**	**0,458**	**1,297**
Hånd	0,906	0,916	0,163	0,588	1,283
Tast	0,879	0,869	0,173	0,458	1,297
Middels	0,848	0,864	0,169	0,458	1,194
Sterk	0,938	0,919	0,156	0,640	1,297
Gutt	0,931	0,934	0,164	0,632	1,297
Jente	0,854	0,853	0,164	0,458	1,263

Figur 11-16 viser at gutter og sterke elever bruker mer preposisjonsfraser enn jenter og middels elever, men at hverken verktøy eller oppgave påvirker antall preposisjoner per klausus. Det er ingen sammenheng med tekstlengde, og bare en moderat korrelasjon mellom håndtekster og tastetekster, $R \approx 0{,}29$.



Figur 11-16: Antall preposisjoner per klausus fordelt etter fire faktorer

11.2.1.4 Variansanalyse

Variansanalysen tar utgangspunkt i den maksimale modellen i (163), med differanse i preposisjonsfrekvens som responsvariabel og de fire dikotome variablene kjønn, skriveferdighet, total tekstlengde og tekstlengdeforskjell som prediktorer, interaksjoner begrenset til 2 nivåer.

(163)
```
lm(posD$prep.kl~(kjønn+ferdighet+lengde+forskjell)^2)
```

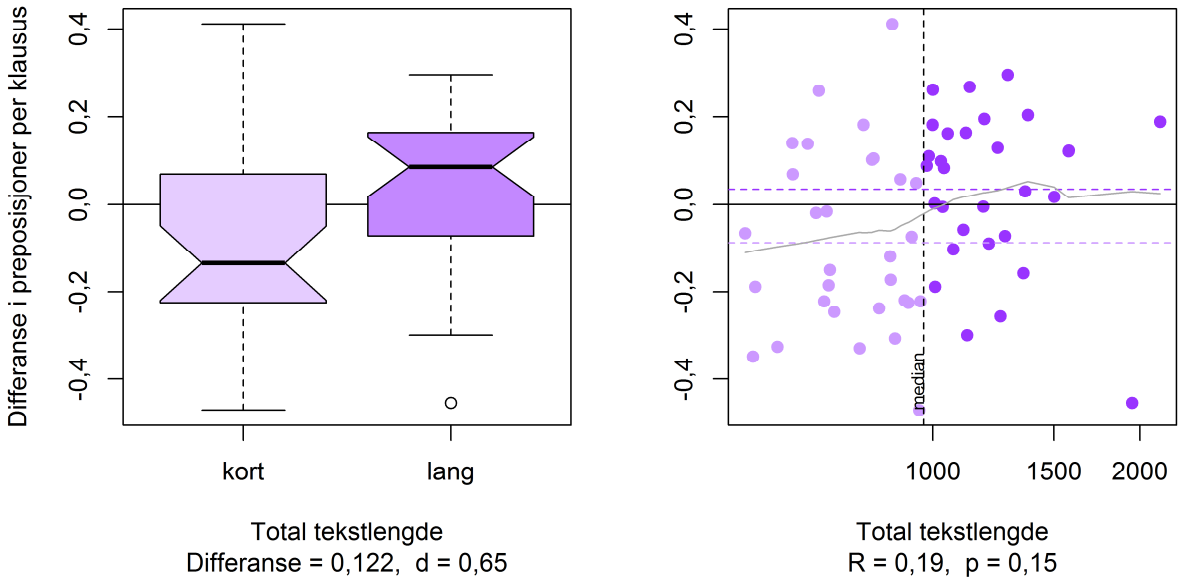
Modellreduksjonen resulterer i en minimal adekvat modell med total tekstlengde som eneste signifikante prediktor, $F \approx 6{,}0$, $p < 0{,}05$.

(164)
```
lm(formula = posD$prep.kl ~ lengde)

             Df Sum Sq Mean Sq F value Pr(>F)
lengde        1 0.2246 0.22460   6.045  0.017 *
Residuals    58 2.1550 0.03716
---
Multiple R-squared:  0.09439,   Adjusted R-squared:  0.07877
```

```
F-statistic: 6.045 on 1 and 58 DF,  p-value: 0.01695
```

`Gvlma` (se 7.2.2.4) rapporterer at premissene for anova-analysen er oppfylt (se appendiks A4). Figur 11-17 viser at elever som generelt skriver langt, bruker flere preposisjonsfraser i tastetekstene, mens det motsatte er tilfellet for elever som generelt skriver kort. Boksdiagrammet viser en klar tendens, d ≈ 0,65. Diagrammet til høyre viser at det ikke er noen tydelig korrelasjon mellom differansen og den totale tekstlengden som kontinuerlig variabel, R ≈ 0,19, og forskjellen mellom kortskrivende og langtskrivende elever virker mindre klar i dette diagrammet.



Figur 11-17: Resultatet av anova-analysen av antall preposisjoner per klausus. Til venstre boksdiagram. Til høyre spredningsdiagram som viser korrelasjonen mellom differanse i preposisjonsfrekvens og total tekstlengde.

11.2.1.5 Oppsummering og diskusjon

Det er ingen absolutt forskjell mellom håndtekster og tastetekster i materialet, men det er en svakt signifikant og substansiell forskjell mellom kortskrivende og langtskrivende elever. De langtskrivende elevene bruker flere preposisjonsfraser i tastetekstene enn i håndtekstene, mens det for de kortskrivende elevene er motsatt.

Ikke mange av variablene i denne undersøkelsen gir utslag for total tekstlengde, og når slike resultater dukker opp, kan man mistenke at det dreier seg om effekt av tekstlengdeforskjell eller skriveferdighet "maskert" som tekstlengde gjennom interaksjon mellom to faktorer. Disse to faktorene samvarierer begge noe med tekstlengde, og det er alltid en risiko for at modellreduksjonen kan fjerne nesten-signifikante effekter av andre faktorer. For denne variabelen er imidlertid dette ikke tilfellet; både ferdighet og tekstlengdeforskjell gir så godt som inget utslag for preposisjoner per klausus, henholdsvis d ≈ 0,004 og d ≈ 0,23, så dette er mest trolig en reell effekt av total tekstlengde.

Det er likevel mulig at total tekstlengde er en variabel som reflekterer en form for ferdighet eller kompetanse, og at dette dreier seg om elever som greier å utnytte tekstbehandlingsverktøyet til å redigere inn mer redigert kompleksitet i klaususene. Siden faktoren i liten grad gir signifikant utslag for andre variabler,[41] er det vanskelig å vite hva slags egenskaper ved tekster eller elever den representerer.

Antall preposisjonsfraser per klausus er altså tenkt å være en indirekte operasjonalisering av antall ledd per klausus, men variabelen er selvfølgelig ikke direkte sammenlignbar med antall ledd, ettersom samme ledd gjerne kan inneholde flere preposisjonsfraser. Men selv om en preposisjonsfrase ikke nødvendigvis svarer til et ledd, vil den normalt uansett tilføre klaususen ekstra kompleksitet, og mengden ekstra kompleksitet er kanskje ikke i særlig stor grad avhengig av om frasen fungerer som et selvstendig ledd eller ikke. I mange tilfeller kan det også være vanskelig å avgjøre om en preposisjonsfrase er et ledd eller en del av et ledd (for eksempel som i (165)), uten at jeg synes det har vesentlige konsekvenser for fortolkningen av denne variabelen.

(165) Ungdommer trenger denne informasjonskilden i studiene i dag. [A1-206]

At preposisjonsfraser kan realisere obligatoriske ledd like vel som adjungerte ledd (som i (166)), eller heller bør analyseres som verbpartikler (som i (167)), har heller ikke vesentlige konsekvenser for fortolkningen av analysen, etter min mening.

(166) Men kan vi legge skylden på dem? [A2-222]
(167) Har du ikke penger, er det bare å møte opp, og smake litt her og litt der. [A2-233]

11.2.2 Adverbiale subklaususer

Antall adverbiale subklaususer per klausus er et annet forsøk på en delvis operasjonalisering av antall adjungerte ledd i klaususen. I likhet med preposisjonsfraser er adverbiale subklaususer en vanlig måte å realisere ikke-obligatoriske ledd på. Selv om preposisjonsfraser og adverbiale subklaususer kan brukes til å uttrykke de samme innholdselementene, bidrar adverbiale subklaususer til en annen type kompleksitet enn preposisjonsfraser.

Som for preposisjoner bruker jeg antall klaususer som målestokk også for adverbiale subklaususer, igjen med den begrunnelsen at variabelen skal være en partiell representasjon av antall ledd per klausus.

11.2.2.1 Hypotese

Biber (1986; 1988) har ikke analysert adverbiale subklaususer generelt, men "*causative subordination*", "*conditional subordination*", "*other adverbial subordinators*" og "*concessive*

[41] Total tekstlengde er en signifikant faktor bare for korte subklaususer (11.1.3).

subordination" separat. Av disse er det bare de to første som spiller en rolle i dimensjon 1, kausal underordning (vekt 0,66) og betingelsesunderordning (vekt 0,32). Årsaksklaususer og betingelsesklaususer er to av de mest frekvente adverbiale klausustyper i elevtekstkorpuset. De har henholdsvis 167 og 282 forekomster av totalt 992 adverbiale klaususer, så de spiller en viktig rolle i variabelen, selv om den mest frekvente adverbialklaususen er den temporale. De andre kategoriene til Biber spiller roller i andre dimensjoner som ikke er så relevante i denne undersøkelsen.

Selv om adverbiale subklaususer på sett og vis har samme funksjon som preposisjonsfraser, å fylle adverbiale, oftest ikke-obligatoriske, ledd i klaususen, og de ved å danne ekstra ledd kan sies å øke kompleksiteten i klaususen, befinner de seg altså i hver sin ende av Bibers dimensjon 1 (1988, s. 102). Også i Hallidays argumentasjon (se 3.1) er det åpenbart at preposisjonsfraser og adverbiale subklaususer representerer hver sin type kompleksitet, og Baron tok også utgangspunkt i adverbiale subklaususer som et muntlig trekk i sin analyse av digital språkbruk (se 3.1). Min hypotese er derfor at adverbiale subklaususer er mest frekvente i tastetekstene.

Med samme argumentasjon som for preposisjoner velger jeg altså også for adverbiale subklaususer klaususen som målestokk; Biber bruker løpeord som målestokk også for subklausus-variablene.

11.2.2.2 Korpussøk og utregning

Adverbiale subklaususer er tagget manuelt i korpuset og er direkte tilgjengelige med søket i (168).

(168) `<clause type="adverbial">`

Som forklart i 4.1.4 er klausustypene kategorisert etter syntaktisk funksjon og ikke form, og det medfører at en del subklaususer som er innledet av det vi normalt ser på som adverbiale subjunksjoner, er kategorisert som nominale, som for eksempel i (169).

(169) Som jeg sa tidligere i dette innlegget, så kan jeg ikke huske når jeg sist leste en bok. [A1-202]

For adverbiale subklaususer er det imidlertid ganske sterkt samsvar mellom de syntaktiske og de formale kriteriene. Av 1813 nominale subklaususer innledes 10 med \fordi\ og 3 med \når\, mens resten innledes på måter som formalt samsvarer med nominale funksjoner. Av 990 adverbiale subklaususer innledes 2 med \at\,[42] mens resten innledes på måter som formalt samsvarer med adverbiale funksjoner.

[42] Ett av disse tilfellene er nok en feiltagging.

Siden hypotesen også for denne variabelen er knyttet til antall per klausus, har jeg regnet ut frekvensen per antall klaususer totalt, som for frekvens av preposisjonsfraser.

$$adverbiale\ subklaususer\ per\ klausus = \frac{adverbiale\ subklaususer}{klaususer}$$

Dette innebærer at antall adverbiale subklaususer inngår både i divisor og dividend i formelen; variabelen blir dermed beslektet med en tetthetsvariabel, og den vil påvirkes av den generelle subklaususfrekvensen og til en viss grad av klaususlengden i korpuset. I 11.1.3.6 argumenterer jeg for en mest mulig teorinøytral og uavhengig målestokk, noe som i dette tilfellet skulle tilsi antall ord. Når imidlertid hypotesen er knyttet til antall adverbiale subklaususer per klausus, er det det logiske å velge klaususen som målestokk. Men tankegangen bak er altså ikke å finne ut hvor stor *andel* av klaususene i korpuset som er adverbiale subklaususer, men å finne ut hvor mange adverbiale subklaususer hver klausus inneholder. Regnestykket og forholdstallet for disse to konseptualiseringene blir riktig nok nært beslektet.

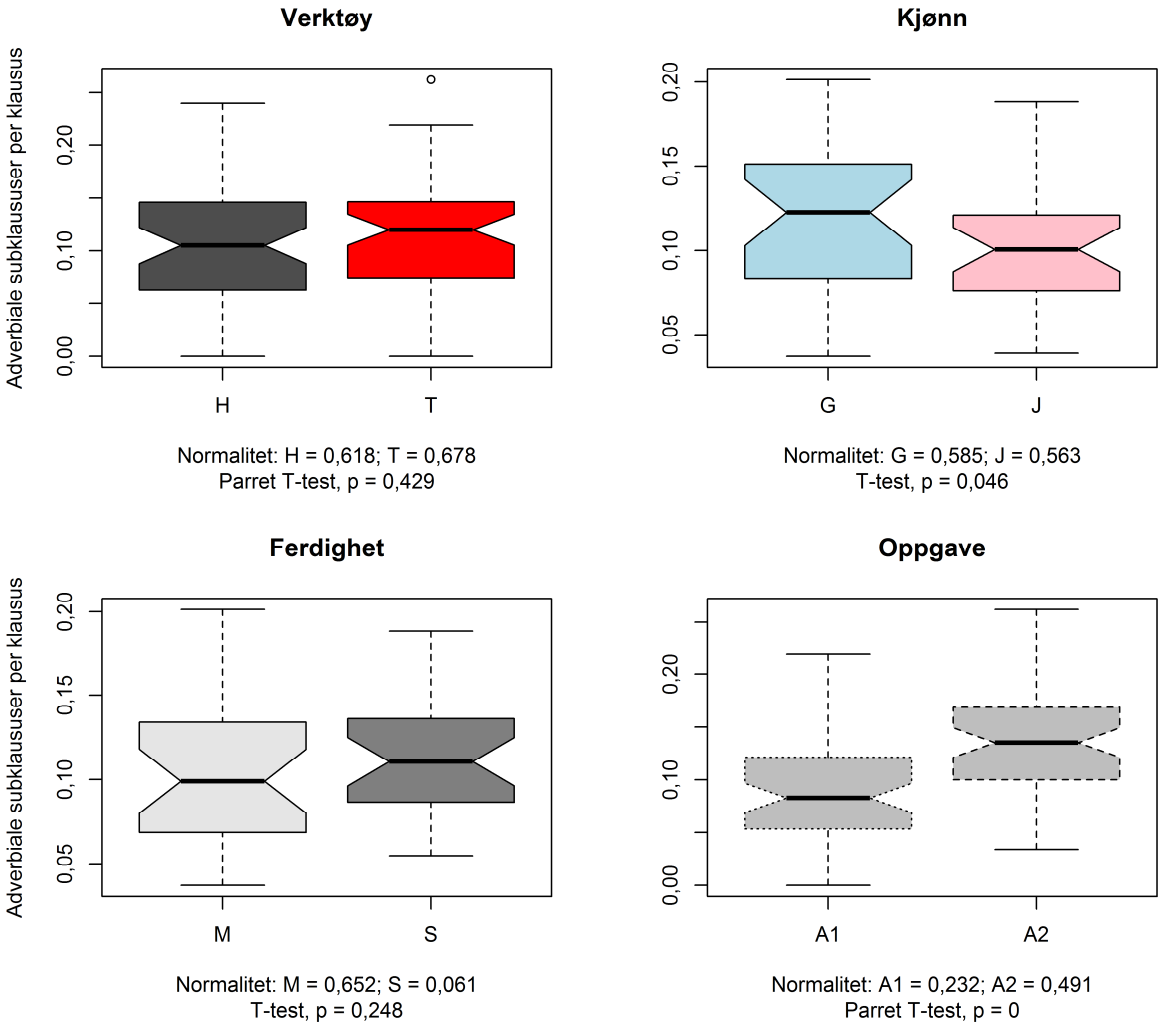
11.2.2.3 Deskriptiv analyse

Tabell 11-5 viser nøkkeltallene for variabelen. Tabellen viser gjennomsnittsverdier på i overkant av 0,1 med standardavvik på noe over 0,05. Selv om nøkkeltallene signaliserer en viss høyreskjevhet, blir utvalget og alle segmentene rapportert som normalfordelte av Shapiro-Wilk-testen. 4 tekster har ingen adverbiale subklaususer; dette gjelder både hånd- og tastetekster, men de er alle "Bøker eller data"-tekster.

Tabell 11-5: Nøkkeltall for antall adverbiale subklaususer per klausus

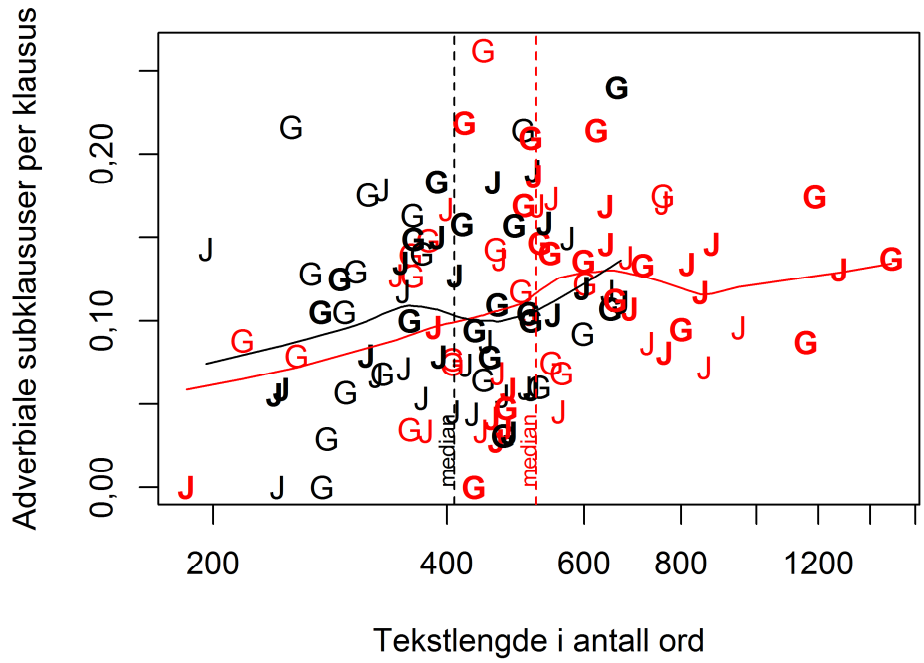
	middelverdi	median	sd	min	maks
Total	**0,110**	**0,108**	**0,055**	**0**	**0,262**
Hånd	0,107	0,105	0,054	0	0,240
Tast	0,114	0,120	0,056	0	0,262
Middels	0,104	0,094	0,055	0	0,262
Sterk	0,116	0,115	0,054	0	0,240
Gutt	0,121	0,120	0,057	0	0,262
Jente	**0,100**	**0,100**	**0,050**	**0**	**0,190**

Figur 11-18 viser at hverken verktøy eller skriveferdighet har påvirkning på adverbiale subklaususer, mens gutter bruker flere enn jenter. Mest markert er overvekten av bruken i "Ungdomsfylla"-tekstene, d ≈ 0,92. Det er *ingen* korrelasjon mellom håndtekster og tastetekster, R ≈ 0,072, så dette synes som en variabel elevene i liten grad har et skrivemønster for, eller at den er svært avhengig av emne eller teksttype.



Figur 11-18: Antall adverbiale subklaususer per klausus fordelt etter fire faktorer

Figur 11-19 viser en viss tendens til korrelasjon med tekstlengde. For tastetekstene er korrelasjonen svak, R ≈ 0,26; for håndteksten er den enda svakere, R ≈ 0,16.



Figur 11-19: Spredningsdiagram for adverbiale subklaususer og tekstlengde. R ≈ 0,26 for tastetekstene, R ≈ 0,16 for håndtekstene.

11.2.2.4 Variansanalyse

Variansanalysen tar utgangspunkt i en maksimal modell med differansen mellom frekvensverdiene som responsvariabel og de fire dikotome faktorene kjønn, skriveferdighet, total tekstlengde og tekstlengdeforskjell som prediktorer, interaksjoner begrenset til 2 nivåer.

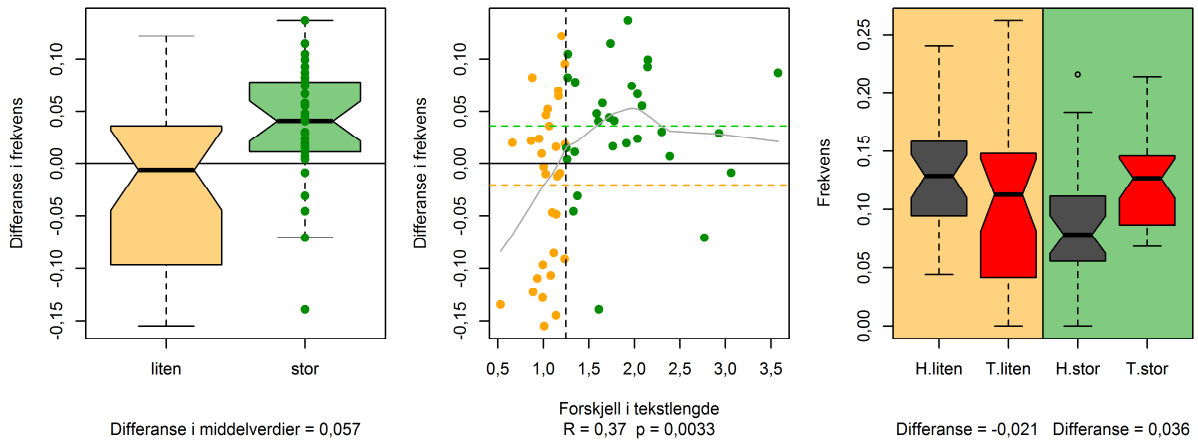
(170) `lm(synD$klaus.adv.kl<-(kjønn+ferdighet+lengde+forskjell)^2)`

Modellen ble redusert til den minimale modellen i (171) med bare tekstlengdeforskjell som signifikant prediktor, F ≈ 10,1, p < 0,01. `Gvlma` (se 7.2.2.4) viser at premissene for anova-analysen er oppfylt (se appendiks A4).

(171)
```
lm(formula = synD$klaus.adv.kl ~ forskjell)

             Df  Sum Sq Mean Sq F value  Pr(>F)
forskjell     1 0.04878 0.04878   10.09 0.00239 **
Residuals    58 0.28053 0.00484
---
Multiple R-squared:  0.1481,    Adjusted R-squared:  0.1334
F-statistic: 10.09 on 1 and 58 DF,  p-value: 0.002394
```

Elever med stor tekstlengdeforskjell bruker flere adverbiale subklaususer i tastetekstene, mens resten av elevene bruker noe færre adverbiale subklaususer i tastetekstene, men effekten er størst for elevene med stor tekstlengdeforskjell. Forskjellen i differansenes middelverdi er D ≈ 0,057, d ≈ 0,83. Figur 11-20 viser også at 25 av 30 elever som skriver mye lengre på tastatur, bruker mer adverbiale subklaususer i tastetekstene (venstre og midtre diagram).



Figur 11-20: Diagrammer som viser resultatet av anova-analysen for antall adverbiale subklaususer per klausus. Til venstre boksdiagram som viser differansene i frekvens. I midten spredningsdiagram som viser korrelasjon med forskjell i tekstlengde. Til høyre boksdiagram som viser frekvenser fordelt etter verktøy og tekstlengdeforskjell.

Figuren viser en sterkere korrelasjon med variabelen tekstlengdeforskjell enn korrelasjonen vi så mellom preposisjonsfrekvens og tekstlengde (11.2.1.4), særlig hvis man ser bort fra de 4 høyeste verdiene for tekstlengdeforskjell (diagrammet i midten). Diagrammet til høyre viser at det særlig er verdiene i håndtekstene som utgjør forskjellen i differanse; elever med stor tekstlengdeforskjell har lavere verdier i håndtekstene. Dette kan ha en viss sammenheng med korrelasjonen med tekstlengde, selv om den er svak.

11.2.2.5 Oppsummering og konklusjon

Tekstlengdeforskjell gir et ganske sterkt utslag for frekvensen av adverbiale subklaususer, og særlig utviser elevene med stor tekstlengdeforskjell forskjell i hvordan de bruker verktøyene. For disse elevene stemmer analyseresultatet med hypotesen, mens den svakere effekten for resten av elevene går i motsatt retning av hypotesen. Det er rimelig å tenke seg at det for den første gruppen er den høyere produksjonshastigheten som påvirker språkbruken i den muntlige, spontane retningen, mens den andre gruppen bruker mer tid med tekstbehandlingsverktøyet på redigering og reformulering, og at deres tastetekster dermed blir mer planlagte og integrerte i strukturen. Imidlertid er det et problem for denne tolkningen at det først og fremst håndtekstene det er forskjell mellom. Det er også mulig at den riktignok svake korrelasjonen med tekstlengde kan ha innvirkning på resultatet.

Effekten for adverbiale subklaususer er vesentlig sterkere enn for den andre operasjonaliseringen av gjennomsnittlig antall ikke-obligatoriske setningsledd. Dette kan nok delvis være en effekt av at preposisjoner er et mer frekvent trekk med mange ulike typer funksjoner, slik Chafe og Danielewicz er inne på, og at dette trekket derfor kan være mindre entydig påvirket av situasjonelle faktorer.

11.2.3 Oppsummering

Jeg har i dette delkapitlet sett på to ulike og på hver sine måter indirekte operasjonaliseringer av antall ledd per klausus. Bakgrunnen for å velge klausus som målestokk i stedet for løpeord er ønsket om at variablene skulle reflektere nettopp antall ledd per klausus, men ettersom ingen andre studier jeg kjenner til, bruker denne målestokken (Biber (1988) bruker løpeord), blir det vanskelig både å sammenligne resultater og å bruke tidligere studier som hypotesedannere.

For preposisjoner er tendensen at elever som skriver lange tekster, putter flere preposisjoner inn i hver klausus i tastetekstene, mens elever som skriver korte tekster, putter færre preposisjoner inn i hver klausus i tastetekstene. For adverbiale subklaususer er tendensen at elever som skriver lengre på tastatur, putter flere adverbiale subklaususer inn i hver klausus i tastetekstene, mens elever som ikke skriver lengre på tastatur, ikke har noen forskjell mellom håndtekster og tastetekster.

Siden preposisjonsfraser og adverbiale subklaususer er to viktige måter å øke antall ledd i en klausus på, kunne det være naturlig å analysere den akkumulerte frekvensen av de to. Imidlertid er preposisjonsfraser så mye mer frekvente enn adverbiale subklaususer at den resulterende variabelen generelt har de samme egenskaper som preposisjonsfrekvens. Pearsons korrelasjonskoeffisient mellom dem er $R \approx 0.95$, og det er lite poeng i å utføre anova-analyse på den summerte variabelen. Et alternativ kunne være å standardisere begge variablene før summering, men ettersom målestokken er per klausus, er det vanskelig å gi en slik variabel noen validitetsmessig fornuftig fortolkning.

Ingen av de to variablene representerer antall ledd per klausus direkte, men de representerer to utbredte måter å øke klaususens lengde på ved å legge til ikke-obligatoriske ledd. Grunnlaget for hypotesene mine antyder at man kan vente seg ulike effekter for de to variablene, men det at resultatet fra anova-analyser blir ulikt for de to variablene, betyr ikke at det er påvist en forskjell mellom variablene. Jeg har ikke utført analyser som kan påvise at variablene er forskjellige, men prinsipalkomponentanalysen i 12.3 viser at de to variablene systematisk dominerer ulike dimensjoner i analysen, noe som tilsier at de faktisk *er* knyttet til ulike egenskaper ved tekstene. Det er heller ingen korrelasjon mellom dem i tekstene ($R \approx -0.03$), noe som også tyder på at de representerer ulike tekstegenskaper.

11.3 Leddenes lengde

Når man ser på klaususlengde som et mål på kompleksitet, finnes det to dimensjoner å øke denne kompleksiteten etter. Den ene dimensjonen er antall ledd, som jeg har presentert i 11.2, og den andre er leddenes lengde, som jeg presenterer i dette delkapitlet.

Heller ikke leddenes lengde er det mulig å trekke ut av korpuset direkte, men jeg har undersøkt tre egenskaper som på hver sin måte har ganske sterk forbindelse med leddlengde, nemlig subklaususfrekvens, ettords forfelt og attributive adjektiver.

Ledd som består av eller inneholder subklaususer, vil normalt ha en sterk tendens til å være lengre enn ledd uten subklaususer, så vi kan nok regne med at subklaususfrekvens korrelerer positivt med gjennomsnittlig leddlengde i en tekst. Forfeltsleddet er det eneste leddet som det er mulig å trekke ut av korpuset med relativt god presisjon, og andelen forfelt som er minimalt korte er derfor en operasjonaliserbar variabel som er beslektet med kort gjennomsnittlig leddlengde. Attributive adjektiver er svært konkret knyttet til gjennomsnittlig leddlengde, ettersom et ledd med et attributivt adjektiv vil være et ord lengre enn det samme leddet uten det attributive adjektivet, i mange tilfeller to ord lengre, ettersom attributivt adjektiv ved bestemthet normalt krever determinativ.

11.3.1 Subklaususfrekvens

Jeg begynner med subklaususfrekvens, som ofte (se 2.3.1) blir sett på som selve det prototypiske grammatiske trekket som representerer syntaktiske kompleksitet.

11.3.1.1 Hypotese

Subklaususfrekvens er i litteraturen først og fremst knyttet til ytringssituasjonen og til skribentens modenhet. Bibers analyse (1988, s. 102) har ikke subklaususfrekvens generelt som en egen parameter, men de parametrene som er knyttet til klaususunderordning i hans første dimensjon, bidrar alle til den enden av aksen som dreier seg om spontanitet eller interaktivitet: *causative subordination* (vekt 0,66), *sentence relatives* (vekt 0,55), *WH clauses* (vekt 0,47) og *conditional subordination* (vekt 0,32). Halliday hevder mer generelt (Halliday & Matthiessen, 2004, s. 654-655) at muntlig (spontan) språkbruk er preget av å være "*grammatically intricate*" ved å bygge opp komplekser med paratakse og hypotakse, men uten å vise til konkret empiri. Halliday skiller imidlertid mellom hypotakse og innføying (Halliday, 1987, s. 73-74) og hevder at bare hypotakse (i tillegg til paratakse) er et muntlig trekk, mens innføying (*embedding*) er et skriftlig trekk. Chafes funn fra 1982 (Chafe, 1982, s. 44) er i hvert fall delvis i tråd med dette når han finner at både *complement clauses* og *relative clauses*[43] er mer frekvente i skriftlige tekster. Han følger dette imidlertid opp senere (Chafe & Danielewicz, 1987, s. 102-103) med en mer generell påstand om at "*more elaborate clausal relations*" er for kognitivt krevende for taleprosessen, men uten å belegge det med empiri. (Hunt, 1965, s. 16-19) viser at subklausustetthet i skriftlige tekster korrelerer positivt med modenhet, noe som kanskje kunne tas til inntekt for at klausal underordning *er* kognitivt krevende, men resultatene hans er ikke direkte anvendbare for min undersøkelse.

Hallidays skille mellom innføying og hypotakse svarer delvis, men bare delvis, til skillet mellom nominalklaususer og restriktive relativklaususer på den ene siden og adverbiale og utdypende relativklaususer på den andre. Visse typer nominalklaususer regner han imidlertid

[43] Det er litt uklart fra Chafes artikkel hvorvidt han har undersøkt *relative clauses* generelt eller bare *restrictive relative clauses*.

som hypotakse. Dessuten skiller ikke elevtekstkorpuset mellom restriktive og utdypende relativklaususer, og det er derfor umulig å bruke Hallidays teori direkte som grunnlag for testbare hypoteser i elevtekstkorpuset slik det er konstruert. Alle Bibers underordninger i dimensjon 1 hører til Hallidays hypotakse, så Bibers resultater er i tråd med Hallidays teorier.

Ut fra dette er det naturlig å hypotetisere en generell tendens til høyere subklaususfrekvens i tastetekstene, samtidig som man kan tenke seg at denne variabelen kan være følsom for redigering med tekstbehandlingsverktøy. I så fall er det et åpent spørsmål i hvilken retning redigeringen vil foregå. Biber og Halliday peker på at skriftlig eller planlagt språk er preget av annen type kompleksitet enn klaususunderordning, men på den annen side kan det godt tenkes at elevene har en oppfatning av sammenhengen mellom underordning og modenhet, slik at de kanskje vil etterstrebe økning av syntaksens kompleksitet i sine redigeringsprosesser. Det er derfor uvisst i hvilken retning elevenes skriveideal går; kanskje er det slik at svakere elever vil streve etter en mer kompleks klausal struktur, mens sterkere elever har sterkere profesjonelle modelltekster og derfor vil streve etter mer informasjonstette klaususer.

Jeg tror skriveverktøyet vil påvirke denne variabelen, men at påvirkningen vil gå i ulike retninger for ulike elever, og at disse effektene kan motvirke hverandre slik at den samlede effekten blir nøytral.

Hunt analyserer altså variabelen subklausustetthet, som er antall subklaususer dividert på antall klaususer totalt. Med utgangspunkt i mitt ønske om å operasjonalisere ulike faktorer som bidrar til t-enhetslengde, mener jeg antall subklaususer per t-enhet (eller hovedklausus) er en mer valid variabel. De to variabelvariantene er monotont kongruente, men valget har konsekvenser for variabelens distribusjon og egenskaper i en statistisk analyse. Jeg velger t-enhet som målestokk og kaller variabelen subklaususfrekvens.

11.3.1.2 Korpussøk

Antall subklaususer kan trekkes rett ut av korpuset ved hjelp av det strukturelle attributtet `<clause>`. Subklaususfrekvens regnes ut som antall subklaususer per t-enhet.

(172) `<clause>`

11.3.1.3 Deskriptiv analyse

Variabelen og alle utvalgene fordelt etter verktøy, kjønn, ferdighet og oppgave er normalfordelt.

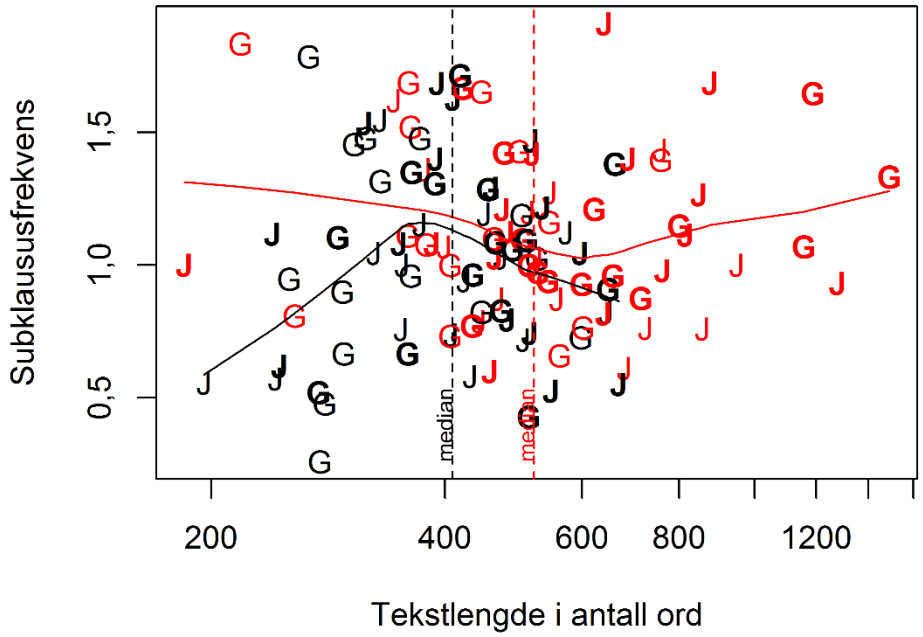
Tabell 11-6: Antall subklaususer per t-enhet

	middelverdi	**median**	**sd**	**min**	**maks**
Total	**1,08**	**1,07**	**0,34**	**0,26**	**1,91**
Hånd	1,03	1,04	0,36	0,26	1,79

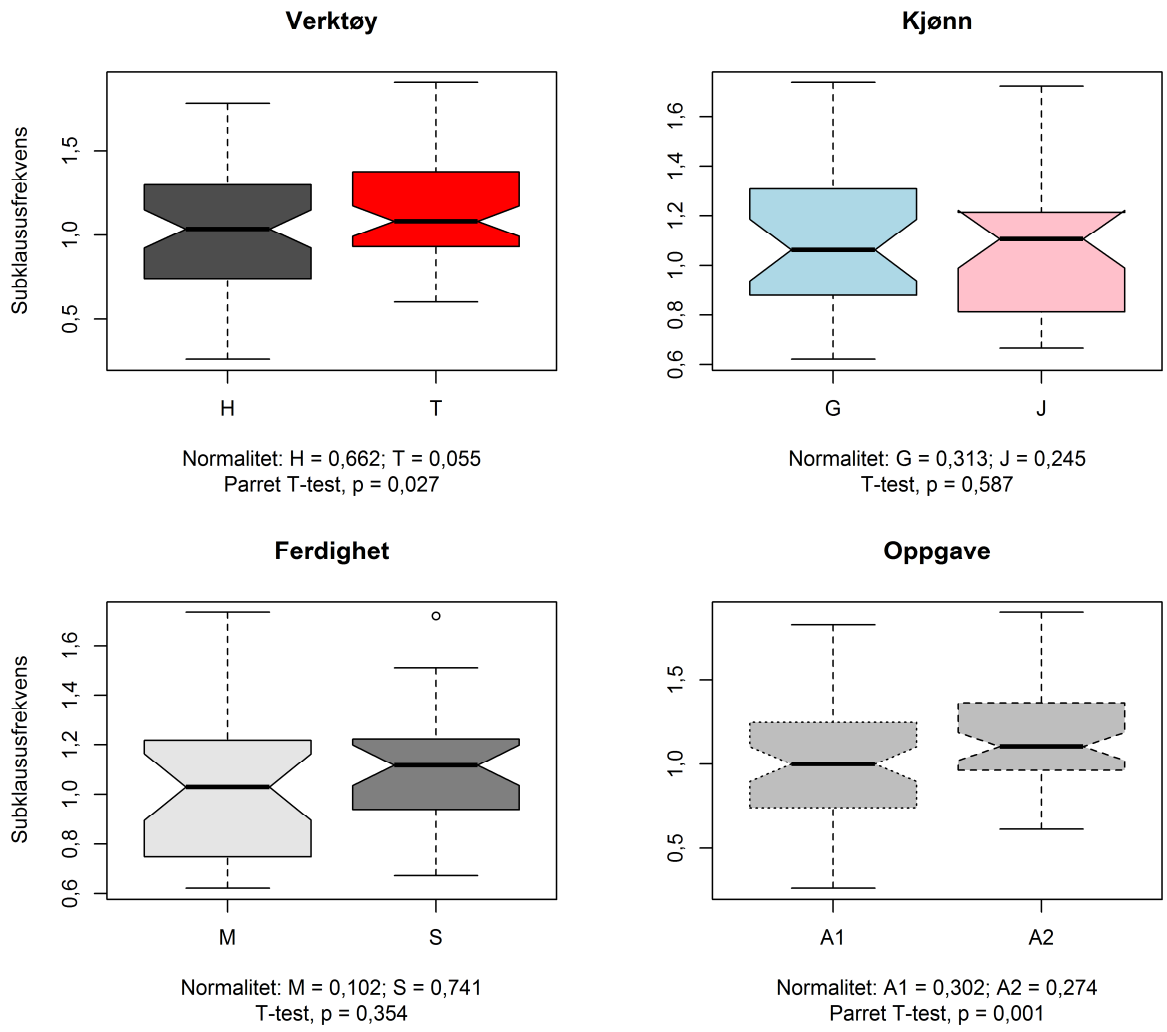
Tast	1,14	1,08	0,32	0,60	1,91
Middels	1,05	1,03	0,35	0,26	1,83
Sterk	1,12	1,09	0,33	0,43	1,91
Gutt	1,10	1,08	0,35	0,26	1,83
Jente	**1,06**	**1,04**	**0,33**	**0,53**	**1,91**

Middelverdien for subklaususfrekvens er ikke langt over 1 (se tabell 11-6); det er altså så vidt over 1 subklausus per t-enhet i gjennomsnitt. Samtidig viser tabellen at standardavviket er på 0,34, og minimums- og maksimumsverdiene viser også at det er stor variasjon, ned mot 1 subklausus for hver fjerde t-enhet, og opp mot 2 subklaususer per t-enhet.

Når det gjelder sammenheng mellom subklaususfrekvens og tekstlengde, viser figur 11-21 at det ikke er noen entydig sammenheng. Som for t-enhetslengde (figur 11-1) kan det se ut til å være en ikke-lineær korrelasjon, og regresjonskurvene fra `lowess` indikerer knekkpunkt i nærheten av medianen for både håndtekster og tastetekster.

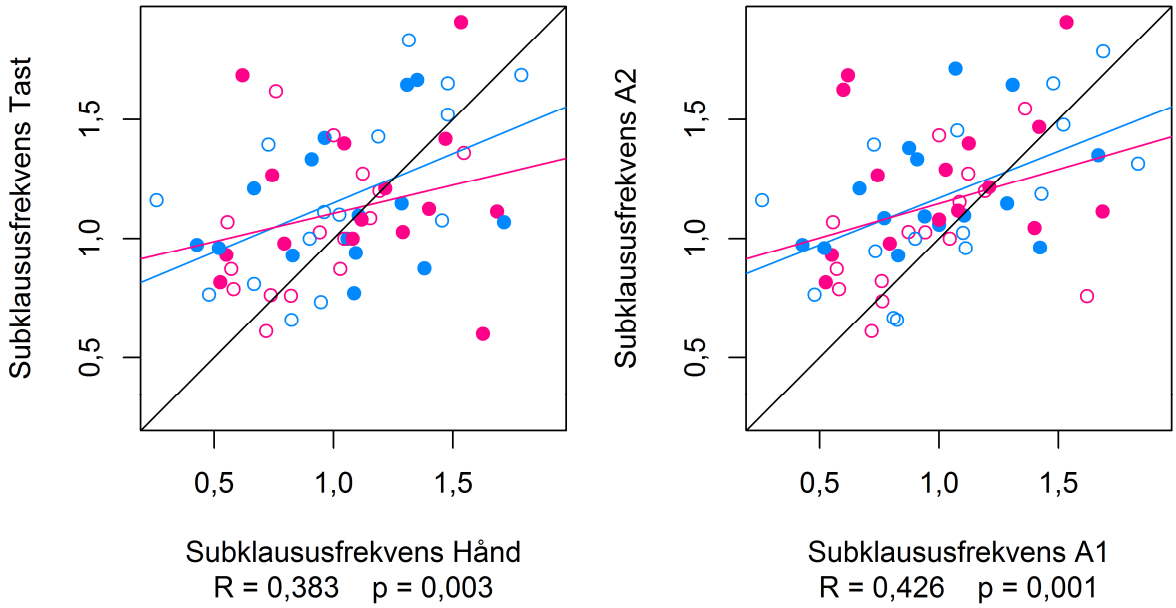


Figur 11-21: Subklaususfrekvens og interaksjon med tekstlengde.



Figur 11-22: Subklaususfrekvens etter verktøy, kjønn, ferdighet og oppgave

Figur 11-22 viser at det er høyest frekvens av subklaususer i "Ungdomsfylla"-tekstene, $d \approx 0.47$, mens det ikke er forskjeller mellom kjønnene og mellom ferdighetsnivåene.



Figur 11-23: Subklaususfrekvens: Korrelasjon mellom verdier i håndtekster og tastetekster (til venstre) og "Bøker eller data" (A1) og "Ungdomsfylla" (A2).

11.3.1.4 Variansanalyse

Variansanalysen tar utgangspunkt i en maksimal modell med differansen mellom frekvensverdiene som responsvariabel og de fire dikotome faktorene kjønn, skriveferdighet, total tekstlengde og tekstlengdeforskjell som prediktorer, interaksjoner begrenset til 2 nivåer.

(173)
```
lm(synD$subkl.te~(kjønn+ferdighet+lengde+forskjell)^2)
```

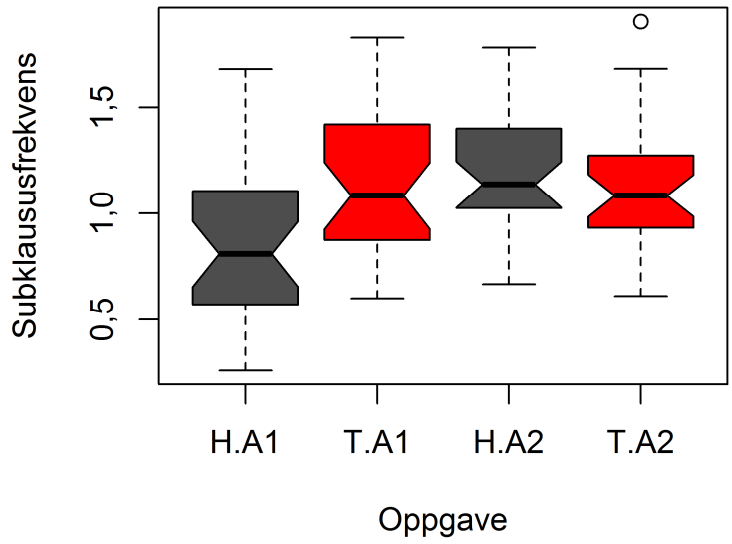
I likhet med for t-enhetslengde resulterer modellreduksjonen i nullmodellen som den minimale adekvate modellen,

(174)
```
lm(formula = synD$subkl.te ~ 1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.11036    0.04866   2.268    0.027 *
---
Residual standard error: 0.3769 on 59 degrees of freedom
```

og denne modellen er svakt signifikant, $t \approx 2{,}27$, $df = 59$, $p < 0{,}05$, $d \approx 0.33$, $D \approx 0.11$. Som figur 11-22 ovenfor viser, er frekvensen av subklaususer høyere i tastetekstene. Shapiro-Wilks normalitetstest rapporterer normalfordeling for differansen av tastetekstverdier og håndtekstverdier ($W \approx 0.985$, $p \approx 0.66$), og for både håndtekster og tastetekster ($W \approx 0.985$, $p \approx 0.66$; $W \approx 0.961$, $p \approx 0.55$), selv om det er en viss høyreskjevhet blant tastetekstene.

Figur 11-23 viser at mye av den forskjellen mellom hånd- og tastetekster som kommer fram i variansanalysen, trolig skriver seg fra de ni håndtekstene som har lavere subklaususfrekvens enn noen av tastetekstene. Disse tekstene er skrevet av både middels og sterke gutter og jenter. De ni elevene står for omtrent halvparten av forskjellen i middelverdier, så det er ikke slik at ulikheten mellom de to segmentene forsvinner helt uten disse ni.



Figur 11-24: Subklaususfrekvens. Interaksjon mellom verktøy og oppgave

Figur 11-24 antyder dessuten interaksjon mellom verktøy og oppgave. De håndskrevne A1-tekstene skiller seg ut med særlig lav subklaususfrekvens, mens det ikke er noen slik forskjell mellom håndskrevne og tastede A2-tekster. Det er mulig å se dette i sammenheng med lærernes inntrykk av at A1-oppgaven fungerte dårligere enn A2-oppgaven, og en lignende interaksjon finnes i gjennomsnittlig t-enhetslengde, mens den speilvendte interaksjonen – altså med *høyere* verdier i håndskrevne A1-tekster – finnes i leksikalsk tetthet og delvis i antall preposisjoner per klausus.

11.3.1.5 Oppsummering og diskusjon

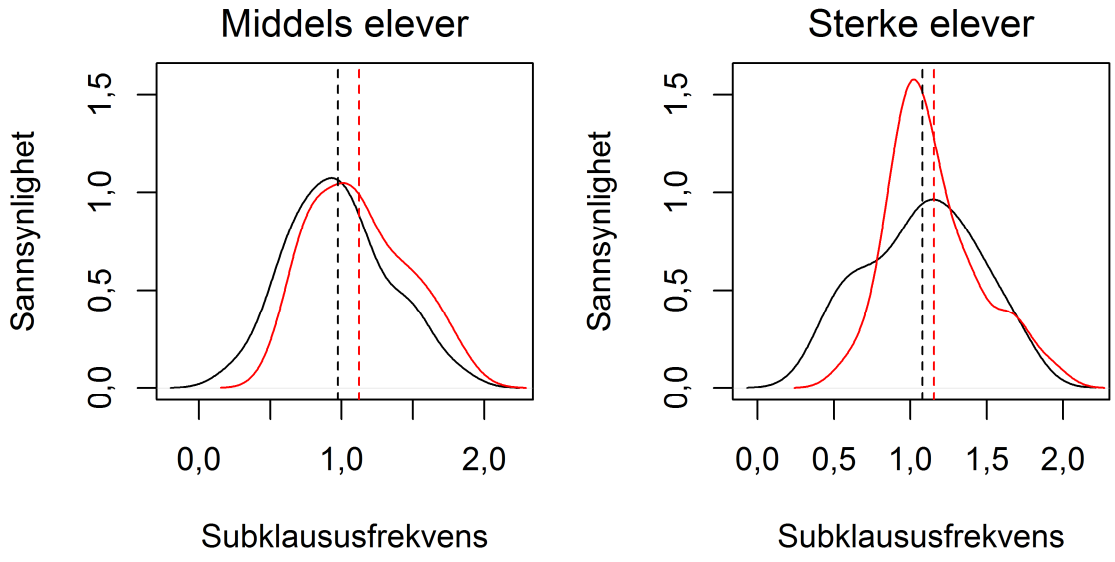
Utgangspunktet for valg av variabelen subklaususfrekvens er sammenhengen mellom antall subklaususer i en t-enhet og t-enhetens lengde. Figur 11-1 og figur 11-21 viser mønstre for t-enhetslengde og subklaususfrekvens som ligner på hverandre. Variansanalysene av de to variablene har også resultert i den samme minimale adekvate modellen, nemlig null-modellen. Figur 10-30 viser at det er en tett sammenheng mellom de to variablene, og Pearsons korrelasjonskoeffisient har en svært høy verdi, $R \approx 0{,}83$. Det virker derfor som en rimelig konklusjon at subklaususfrekvens spiller en viktig rolle i t-enhetslengden, slik Hunt pekte på i 1965 (s. 37).

Når det gjelder forskjell mellom hånd- og tastetekster, viser analysen generelt at det er noe høyere frekvens av subklaususer i tastetekstene, noe som er i tråd med den mest grovformulerte hypotesen i 11.3.1.1 og bekrefter tastetekster som en noe mer spontant og "muntlig" teksttype.

Med hensyn til delhypotesen om virkningen av redigering i tekstbehandlingsverktøyet avslører ikke variansanalysen noen slik effekt. Jeg vil forfølge dette spørsmålet noe mer, men understreker at de tendenser som viser seg, ikke støttes av variansanalysen, og at det dermed hefter en god del usikkerhet rundt realiteten i disse tendensene. Én av årsakene til at

tendensen ikke dukker opp i variansanalysen, kan rett og slett være at en parameter som trolig er ganske avgjørende her, nemlig subklaususfrekvens i håndtekstene isolert, ikke var en parameter i variansanalysen.

Det er kjent (se 3.3.2) at det først og fremst er sterke elever som redigerer tekst etter at den er festet på papir eller skjerm. I dette eksperimentet er tidsrammen for skrivingen ganske knapp, og det er derfor en plausibel hypotese at det i størst grad er med tekstbehandlingsverktøy at de sterke elevene greier å redigere. I så fall kan det være nyttig å se etter effekten av verktøyet isolert i segmentet av sterke elever. Hvilken retning redigeringen forskyver subklaususfrekvensen i, er imidlertid mer usikkert, ettersom vi har to motstridende delhypoteser omkring dette. Figur 11-25 indikerer at middels elever har en generell økning i subklaususfrekvens i tastetekstene; dette er i tråd med både den generelle hypotesen knyttet til produksjonshastighet og den spesielle hypotesen knyttet til at middels elever vil redigere teksten i retning av større klausal kompleksitet.



Figur 11-25: Tetthetskurver over subklaususfrekvens for henholdsvis middels og sterke elever. Svarte kurver representerer håndtekster; røde kurver representerer tastetekster. Stiplede linjer viser middelverdier.

Figuren viser imidlertid en annen tendens for sterke elever, der det ikke er noen like tydelig endring i middelverdi. Det ser derimot ut til at elever med lave verdier i håndtekstene får en økning, mens elever med høye verdier i håndtekstene får en reduksjon. Dette resulterer i en tydelig tendens til lavere varians i tastetekstene for de sterke elevene. Hverken en Kolmogorov-Smirnov-test ($D = 0.2$, $p \approx 0.59$) eller en F-test ($F \approx 1.49$, $p \approx 0.29$) gir imidlertid signifikant resultat med hensyn til distribusjonen mellom de sterke elevenes hånd- og tastetekster.

Tetthetskurven antyder kanskje at sterke elever som skriver med lav subklaususfrekvens for hånd, forskyver oppover (mot modenhetsmålet), mens sterke elever som skriver med høy

frekvens for hånd, forskyver nedover (mot det skriftlige idealet).[44] Om vi ser på de enkelte individenes verdier, finner vi støtte for denne tolkningen; de 12 sterke elevene med lavest verdier for hånd har alle sammen høyere verdier i tastetekstene, mens det motsatte gjelder for 15 av de 18 elevene med høyest verdier for hånd. En slik tendens *kan* kanskje forklares med at det finnes et stilideal som de sterke elevene bruker tekstbehandlingsverktøyet til å forsøke å nærme seg.

Bibers variant av variabelen, antall subklaususer per antall løpeord, resulterer i en ikke-signifikant nullmodell, $t \approx 1{,}62$, $p \approx 0{,}11$, med omtrent de samme tendensene som over, men noe lavere verdier i tastetekstene enn i håndtekstene for de sterke elevene. Shapiro-Wilks normalitetstest rapporterer om normalitet i både håndtekstutvalget og tastetekstutvalget, $W \approx 0{,}991$, $p \approx 0{,}92$; $W \approx 0{,}962$, $p \approx 0{,}058$, men med noe høyreskjevhet blant tastetekstene.

11.3.2 T-enheter med ett ord i forfelt

Det eneste ikke-verbale leddet som er mulig å trekke automatisk ut av korpuset med ganske stor nøyaktighet, er forfeltsleddet. Forfeltet representerer selvfølgelig bare ledd i én posisjon, og det er dermed ikke representativt for alle ikke-verbale ledd. De fleste syntaktiske funksjoner kan imidlertid opptre i forfeltet, og slik sett representerer forfeltet potensielt mange ledd- og frasetyper i tekstene. Dessuten er gjerne forfeltets egenskaper forbundet med tekstlige egenskaper mer generelt. Når omfangsrike ledd står i forfeltet, kan dette ofte bidra til å gjøre t-enheten "tung" og komputasjonelt kompleks, jf. det såkalte vektprinsippet (L. A. Kulbrandstad, 2005, s. 262).

Ingen undersøkelser av stil eller register som jeg kjenner til, har brukt forfeltslengde i kvantitative stilistiske analyser. Basert på hvordan forfeltslengde er vurdert som stilmarkør, sammen med argumentasjonen og resultatene fra tidligere registeranalyser, er det likevel rimelig å anta at korte forfelt er et spontant trekk. Jeg har derfor som hypotese at tastetekster har kortere forfelt enn håndtekster. Å trekke ut det nøyaktige antall ord i hvert forfelt har imidlertid vist seg krevende rent regnekraftmessig for korpusmaskinen, og jeg har derfor forenklet variabelen til andel t-enheter med akkurat ett ord i forfeltet. Hypotesen er dermed at tastetekstene har en høyere andel t-enheter med akkurat ett ord i forfeltet enn håndtekstene.

Gode skrivere kan imidlertid bruke tekstbehandlingsverktøyet til å gjøre teksten mer integrert i stilen, og forlenging av forfeltet er en slik skriftliggjøring som man kan tenke seg er relativt enkelt å oppnå gjennom redigering i tekstbehandlingsprogram. En delhypotese er dermed at visse elevsegmenter kan ha lavere andel t-enheter med akkurat ett ord i forfeltet.

---

[44] Denne formuleringen antyder at håndtekstene er skrevet først, og at elevene deretter forsøker å endre skrivemønsteret sitt med pc-verktøyet. Dette er ikke tilfellet; halvparten av elevene har skrevet tasteteksten sin først (6.2).

Det er nok likevel lite trolig at strekking av forfeltet er noe elever i VG1 bedriver bevisst; dette må i så fall være en følge av mer generell redigering som går ut på å utvide ledd.

11.3.2.1 Korpussøket

Det er ganske enkelt å finne alle t-enheter med nøyaktig ett ord i forfelt, i tillegg til en eventuell konjunksjon i forbinderfeltet, som vist i (175). Ettersom \så\ ikke er tagget som konjunksjon i korpuset, må disse konjunksjonene søkes etter spesielt, som vist i (176); den manuelle taggingen av subklaususer utelukker \så\ som subjunksjon.

(175)
```
<t-unit>[<>]*[features=("konj")]?([<>] | [lemma="\$.*"])*[!<> &
lemma!="\$.*" & features=(!"konj")][<>]*[features=("@fv") &
path=("t-unit" !"clause")][!</t-unit>]*</t-unit>
```

(176)
```
[<t-unit> & attributes- != "type=(.*)"] ([<> & !<clause> &
!</clause>]|[lemma="\$.*"])* [word="så" %c] ([<> & !<clause> &
!</clause>]|[lemma="\$.*"])* [features=(!"@fv") & !<>] ([<> &
!<clause> & !</clause>]|[lemma="\$.*"])* [features=("@fv")]
```

Antall t-enheter er den åpenbare målestokken for denne variabelen, men jeg begrenset analysen til indikative t-enheter, ettersom lette eller manglende forfelt er det normale i interrogative og imperative t-enheter. Også finitte fragmenter ble utelatt ettersom mange av disse har elidert subjekt og tomt forfelt, mens mange andre mangler finitt verbal i hovedklausus; ingen finitte fragmenter i korpuset har akkurat ett ord i forfelt.

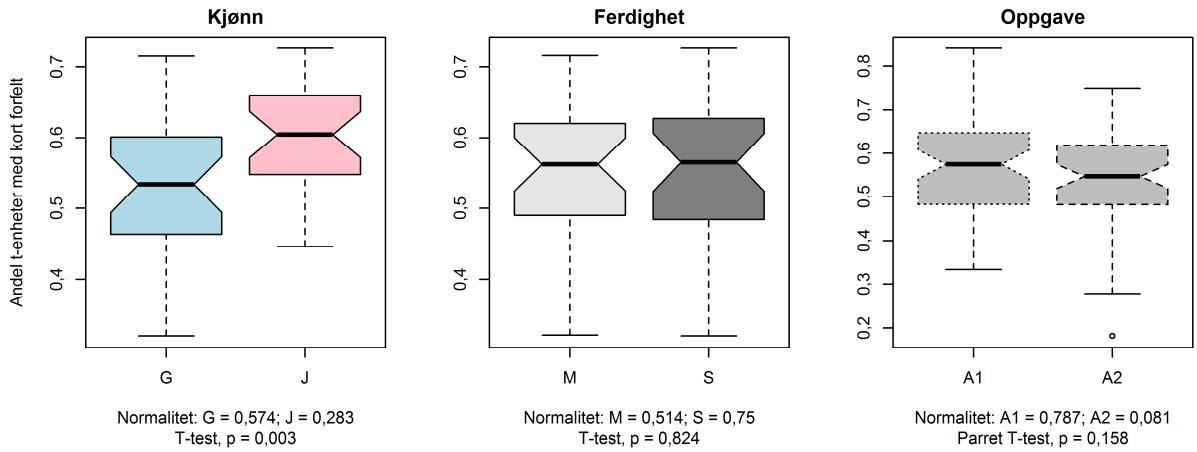
11.3.2.2 Deskriptiv analyse

I beskrivelsene av variabelen vil jeg i det følgende bruke betegnelsen "andel t-enhet med kort forfelt". Tabell 11-7 viser nøkkeltallene.

Tabell 11-7: Nøkkeltall for andel t-enheter med kort forfelt

	middel	median	sd	min	maks
Total	**0,561**	**0,559**	**0,114**	**0,182**	**0,840**
Hånd	0,570	0,565	0,122	0,182	0,792
Tast	0,552	0,556	0,106	0,333	0,840
Middels	0,558	0,560	0,117	0,182	0,840
Sterk	0,564	0,552	0,112	0,278	0,792
Gutt	0,526	0,525	0,116	0,182	0,792
Jente	**0,596**	**0,611**	**0,101**	**0,348**	**0,840**

Tabellen viser at også denne variabelen har stor spredning, fra under hver femte t-enhet med kort forfelt til godt over 4 av 5 t-enheter med kort forfelt. Begge ytterpunktene er skrevet av middels elever. Middelverdiene ligger i overkant av halvparten t-enheter med kort forfelt, med standardavvik i overkant av 0,1.



Figur 11-26: Andel t-enheter med kort forfelt etter kjønn, ferdighet og oppgave

Figur 11-26 viser at ferdighet og oppgave ikke har noen innvirkning på variabelen, men jenter har en vesentlig høyere andel t-enheter med korte forfelt enn gutter, $d \approx 0{,}80$ (logit-transformerte verdier, gjennomsnitt for hver elev), $D \approx 0{,}07$ (nominelle verdier for enkelttekster). Det er ingen korrelasjon mellom andel t-enheter med korte forfelt (logit-transformerte) og tekstlengde (log-transformerte), $R \approx -0{,}06$. Korrelasjonen mellom håndtekster og tastetekster (logit-transformerte) er middels til svak, $R \approx 0{,}30$, så dette er antagelig en variabel med stor variasjon også intraindividuelt.

11.3.2.3 Variansanalyse

Ettersom denne variabelen har typiske forholdstallegenskaper og verdier i nærheten av både 0 og 1, har jeg brukt logit-transformasjon av variabelen i variansanalysen. Responsvariabelen for variansanalysen er dermed differansen mellom de logit-transformerte forholdstallene for taste- og håndtekster. Prediktorvariablene i den maksimale modellen er de fire dikotome faktorene kjønn, skriveferdighet, total tekstlengde og tekstlengdeforskjell, interaksjoner begrenset til 2 nivåer.

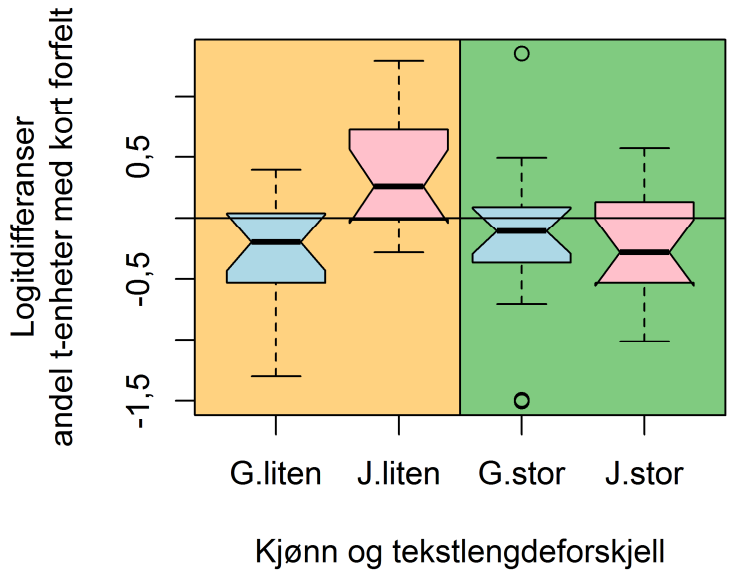
(177)
```
lm(synD$logit.TEind.1ffF~(kjønn+ferdighet+lengde+forskjell)^2)
```

Den maksimale modellen i (177) blir redusert til den minimale adekvate modellen i (178), $F \approx 4{,}46$, $p < 0{,}01$.

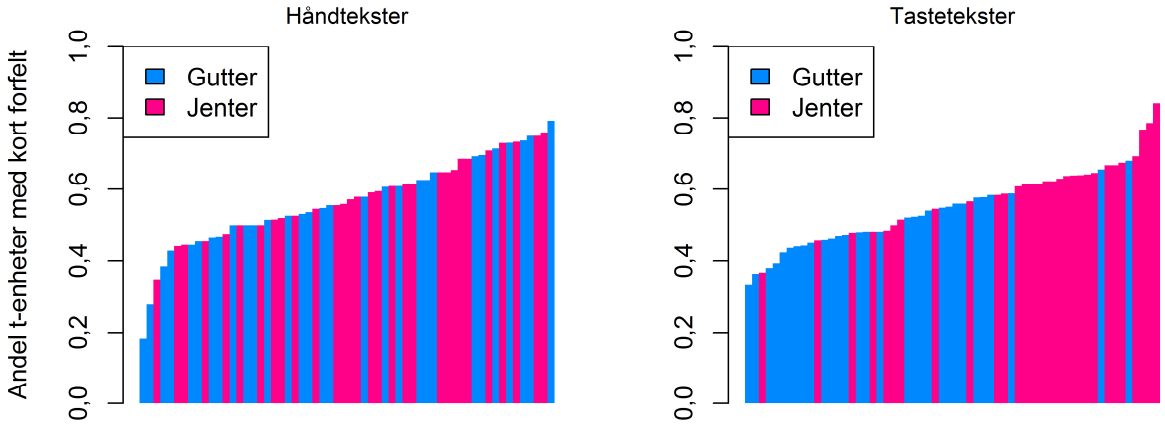
(178)
```
lm(formula = synD$logit.TEind.1ffF ~ kjønn + forskjell +
kjønn:forskjell)
                Df Sum Sq Mean Sq F value Pr(>F)

kjønn            1  1.568  1.5679   5.351 0.0244 *
forskjell        1  0.940  0.9398   3.207 0.0787 .
kjønn:forskjell  1  1.410  1.4098   4.811 0.0324 *
Residuals       56 16.409  0.2930
---
Multiple R-squared:  0.1927,    Adjusted R-squared:  0.1495
F-statistic: 4.456 on 3 and 56 DF,  p-value: 0.007069
```

Premissene for variansanalysen er oppfylt (se appendiks A4), ifølge `gvlma` (se 7.2.2.4). Det er ingen generell effekt på responsvariabelen. Figur 11-27 viser interaksjonen mellom de gjenværende prediktorene i modellen. Figuren antyder en tendens til at andelen t-enheter med ett ord i forfelt er lavere i tastetekster, men at segmentet av jenter med liten forskjell i tekstlengde motvirker en helhetlig tendens og et generelt resultat. Tukeys HSD-test på den minimale modellen viser at segmentet av jenter med liten tekstlengdeforskjell er signifikant forskjellig fra hvert av de tre andre segmentene, $p < 0{,}05$ i hvert tilfelle. Det er interessant å merke seg interaksjonen her; det er altså forskjell på jenter med stor og liten lengdeforskjell ($d \approx 1{,}18$), men ingen slik forskjell blant guttene. Det er også viktig å merke seg at selv om anova viser generell forskjell på gutter og jenter, er det altså bare et definert segment av jentene som forårsaker denne forskjellen.



Figur 11-27: Logitdifferanser i andel indikative t-enheter med kort forfelt. Figuren viser den signifikante interaksjonen mellom kjønn og forskjell i tekstlengde.



Figur 11-28: Andel indikative t-enheter med kort forfelt. Fordeling etter kjønn i håndtekster til venstre og tastetekster til høyre.

Kjønn er altså parameteren med sterkest innvirkning på denne variabelen, og gutter har lavere middelverdi enn jenter. Skreddiagrammet i figur 11-28 illustrerer imidlertid i hvor sterk grad kjønnsforskjellen gjelder bare tastetekstene; skreddiagrammet til venstre viser at fordelingen av kjønn i håndtekstene er ganske lik, mens den skjeve fordelingen blant tastetekstene i diagrammet til høyre er påfallende. Figuren viser at det er sterk konsentrasjon av kjønn i begge ender av verdiomfanget for tastetekstene; det er en overrepresentasjon av gutter som har lave verdier, og en enda sterkere overrepresentasjon av jenter som har høye verdier.

11.3.2.4 Oppsummering og diskusjon

Tendensen blant majoriteten av elevene er at andel t-enheter med bare ett ord i forfeltet er *lavere* i tastetekstene enn i håndtekstene, altså det motsatte av den hypotesen som ble fremsatt i begynnelsen av avsnittet. Dette skulle tyde på at for denne majoriteten er det redigeringsmulighetene som er den sterkeste faktoren når det gjelder akkurat denne variabelen. Det er imidlertid inget stort flertall av elever dette gjelder, kun 35 av 60, og resultatet for gruppen som helhet er ikke signifikant. For de fleste elevene er dessuten forskjellene små.

Det sterkeste resultatet er at segmentet av jenter med liten forskjell i tekstlengde mellom hånd- og tastetekster skiller seg klart både fra guttene og fra resten av jentene ved at de har en høyere andel t-enheter med ett ord i forfelt i tastetekstene. Resten av utvalget er temmelig homogent. Det er viktig å merke seg at det – til tross for resultatet av variansanalysen – *ikke* finnes generelle forskjeller mellom gutter og jenter. Det vi har funnet, er et segment av jenter som som gruppe oppfører seg annerledes enn flertallet ved at de har den hypotetiserte forskyvningen av variabelen i retning av høyere verdier i tastetekstene. Vi har også funnet at flertallet av de som har størst negativ differanse, er gutter. Det som ellers kjennetegner analysen, er at parameteren som skiller jentene i to grupper, ikke har noen innvirkning på guttene.

Resultatet kan sees i sammenheng med andre variabler som viser at gutter i større grad enn jenter synes å fylle inn mer materiale i klaususene sine. Klaususlengde, preposisjoner, adverbiale subklaususer og attributive adjektiver er alle slike eksempler, og ordlengde, leksikalsk ordlengde og MOSTTR peker også i retning av at gutter skriver mer planlagt eller redigert. Dessuten ser vi også i flere variabler at tekstbehandlingsverktøyet hjelper dem til å oppnå dette. Dette gjelder først og fremst preposisjoner og de nevnte leksikalske variablene. Et mindretall av jentene ser ut til å bruke verktøyet til å gjøre tastetekstene mer spontane eller "muntlige" i formen.

Når det gjelder størrelsen på effekten for denne variabelen, gjør logit-transformasjonen den vanskelig å tolke direkte. *Differansen* av logit-verdier kan ikke tilbakeføres til forholdstall mellom 0 og 1. Imidlertid illustrerer skreddiagrammet i figur 11-28 den *rangeringsmessige*

forskjellen mellom kjønnene svært tydelig, og verdien av Cohens *d* mellom de to jentesegmentene er også den sterkeste effekten jeg har målt i hele materialet.[45]

11.3.3 Attributive adjektiver

Attributive adjektiver fører nærmest per definisjon til lengre ledd, ettersom adjektivet kommer i tillegg til en frasekjerne. (Substantivfrase uten substantivkjerne, som \de eldre\ er såpass lite frekvent at de spiller liten rolle i en kvantitativ undersøkelse.) Når frasen er i bestemt form, må dessuten adjektivet normalt følges av et determinativ, slik at det fører til en økning i fraselengde på to ord.

Hos Biber (1988, s. 102-103) er attributive adjektiver en salient faktor i dimensjon 1, noe som gjør at jeg hypotetiserer lavere frekvens av attributive adjektiver i tastetekster enn i håndtekster. Biber undersøker ikke frekvens av adjektiver generelt og finner ingen saliente effekter av predikative adjektiver.

11.3.3.1 Korpussøk

Attributive adjektiver er eksplisitt annotert i korpuset og kan gjenfinnes av følgende søkestreng:

(179) `[features='.* @adj> .*']`

Imidlertid blir en rekke bestemmerledd klassifisert av taggeren som attributive adjektiver, for eksempel \mange\ og \mye\, slik jeg peker på i diskusjonen om leksikalsk tetthet i 9.2.1. Alle forekomster av \mange, \mye\ og ordenstall har jeg fjernet. Dessuten har jeg manuelt fjernet 7 forekomster av \faktisk\, 19 forekomster av \få\, 12 forekomster av \lite\ brukt som determinativ, \masse\, samt 2 andre feiltagginger.

Selv om denne variabelen er tett knyttet til ordklassen adjektiv, er definisjonen av attributive adjektiver tettere knyttet til en spesifikk syntaktisk distribusjon. Det er derfor et naturlig spørsmål om målestokken for attributive adjektiver bør forbedres fra per løpeord til en målestokk som er knyttet til antall potensielle omgivelser. Hvert substantiv er en potensiell omgivelse for ett eller flere attributivt adjektiv, og en mer relevant målestokk kunne derfor være per substantiv, eller eventuelt per klausus, for å unngå uheldige interaksjoner med substantivfrekvens. For å unngå interaksjoner og beholde sammenlignbarheten både med andre variabler og med Bibers analyser, bruker jeg løpeord som målestokk.

---

[45] Effekten for det entropiske målet for leksikalsk variasjon mellom gutter med liten tekstlengdeforskjell og jenter med liten tekstlengdeforskjell er større, d ≈ 1,23 (se 10.5.1.2), men jeg har vurdert denne variabelen som lite valid (10.5.1.3), og den er ikke med i den videre analysen.

11.3.3.2 Deskriptiv analyse

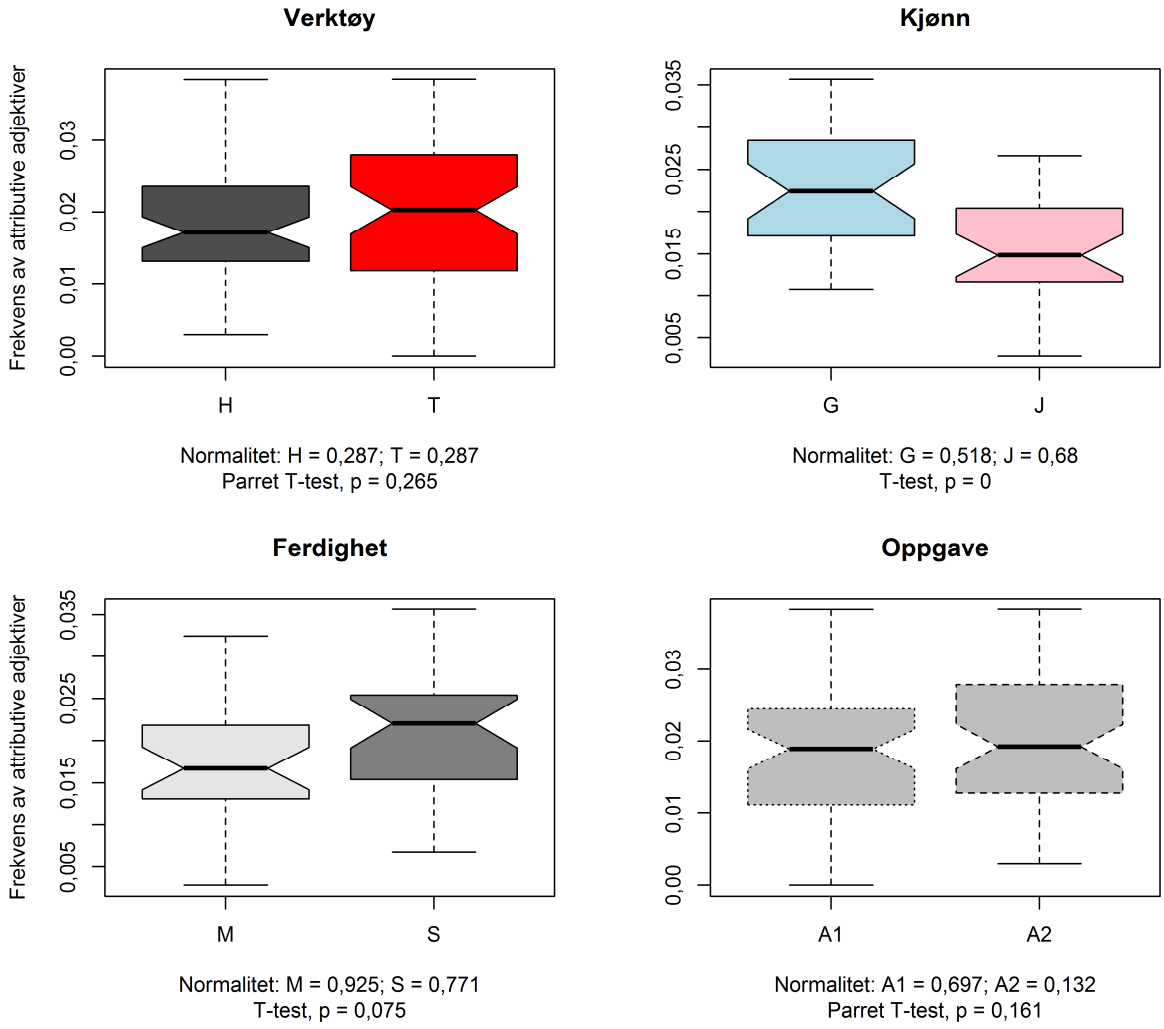
Tabell 11-8 viser at attributive adjektiver ikke er noe veldig frekvent trekk i tekstene. Gjennomsnittet ligger på i underkant av 2 per 100 ord, med spredning fra 0 til bortimot 4 per 100 ord. 1 tekst har ingen attributive adjektiver. Dette er en tekst som er skrevet på tastatur av ei sterk jente, og det er ikke en særlig kort tekst, slik man kanskje kunne ha trodd, men en ganske gjennomsnittlig lang tekst på 460 ord.

Tabell 11-8: Attributive adjektiver, frekvens per løpeord

	middelverdi	**median**	**sd**	**min**	**maks**
Total	**0,0191**	**0,0190**	**0,0089**	**0**	**0,0384**
Hånd	0,0184	0,0172	0,0078	0,0030	0,0384
Tast	0,0198	0,0203	0,0098	0	0,0384
Middels	0,0174	0,0170	0,0085	0,0027	0,0371
Sterk	0,0208	0,0207	0,0090	0	0,0384
Gutt	0,0225	0,0230	0,0087	0,0043	0,0384
Jente	0,0158	0,0152	0,0078	0	0,0329

De relevante utvalgene er normalfordelte, til tross for at man kunne vente en viss gulveffekt ettersom verdiene altså ligger ganske nær 0. Det er ingen nevneverdig korrelasjon med tekstlengde, $R \approx 0{,}16$, mens korrelasjonen mellom håndtekstene og tastetekstene er middels sterk, $R \approx 0{,}39$.

Boksdiagrammene i figur 11-29 nedenfor viser at kjønn er den eneste faktoren som skiller på denne variabelen; gutter har høyere frekvens enn jenter. Effekten er ganske sterk, $d \approx 1{,}03$.



Figur 11-29: Attributive adjektiver, frekvens per løpeord

11.3.3.3 Variansanalyse

Variansanalysen tar utgangspunkt i en maksimal modell med differansen mellom frekvensverdiene som responsvariabel og de fire dikotome faktorene kjønn, skriveferdighet, total tekstlengde og tekstlengdeforskjell som prediktorer, interaksjoner begrenset til 2 nivåer.

(180)
```
lm(posD$AA.redF ~ (kjønn + ferdighet + lengde + forskjell)^2)
```

Med utgangspunkt i den maksimale modellen i (180) resulterte modellreduksjonen i en nullmodell uten signifikant effekt, $t \approx 1{,}13$, $p \approx 0.26$. Begge delutvalgene er normalfordelte, $W \approx 0{,}98$, $p \approx 0{,}29$; $W \approx 0{,}98$, $p \approx 0{,}29$. Det er heller ingen interaksjon mellom skriveverktøy og oppgave.

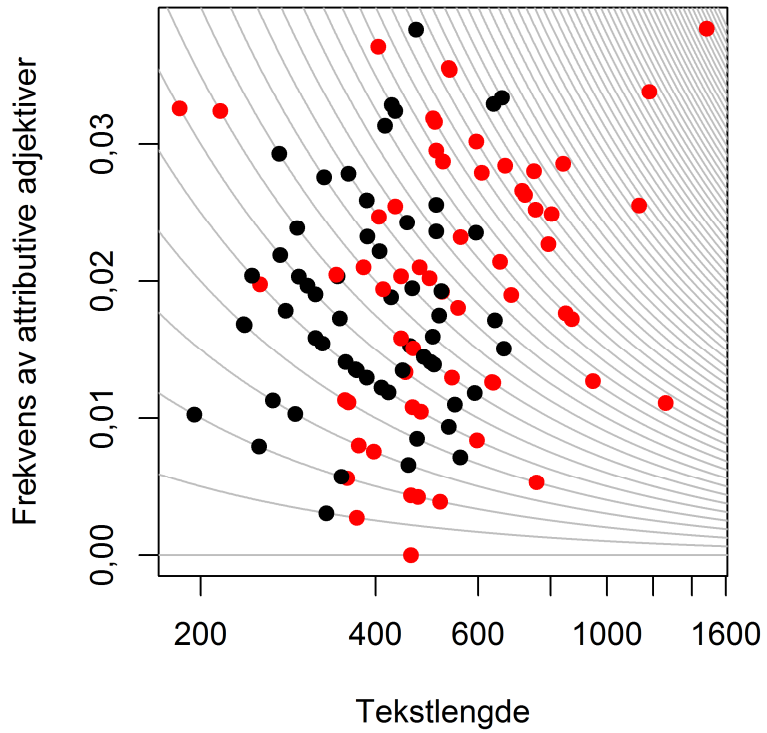
(181)
```
lm(formula = posD$AA.redF ~ 1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001436   0.001275   1.126    0.265
---
Residual standard error: 0.009877 on 59 degrees of freedom
```

11.3.3.4 Oppsummering og diskusjon

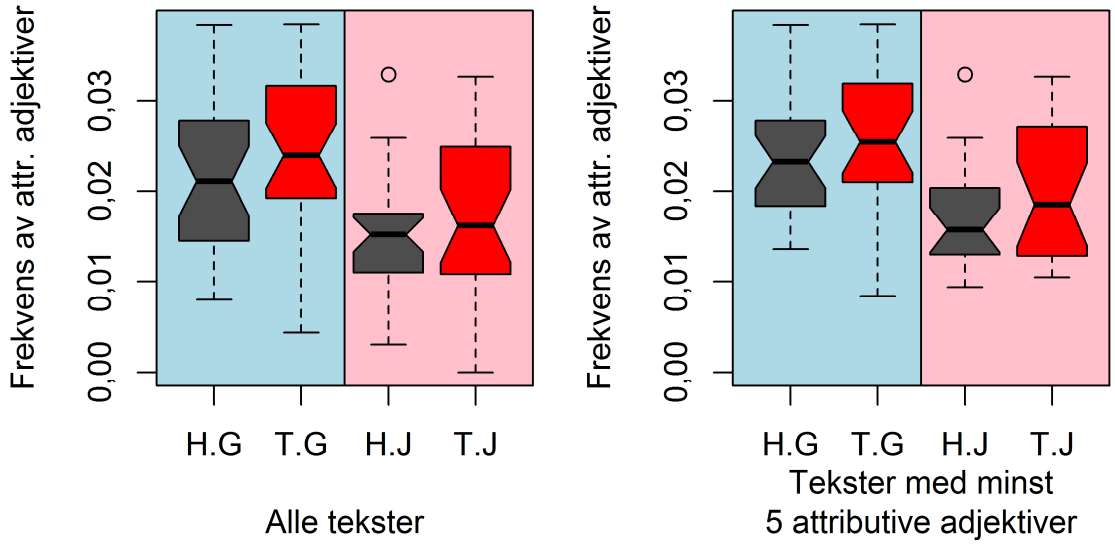
Det er vesentlig forskjell på gutter og jenter i hvor mye attributive adjektiver de bruker, men skriveverktøy har ingen signifikant innvirkning på frekvens av attributive adjektiver, hverken på gruppenivå eller på noen av enkeltsegmentene.

En mulig delforklaring på at variansanalyse ikke gir signifikante resultater, kan være de lave måleverdiene, jf. spredningsdiagrammet i figur 11-30 nedenfor.



Figur 11-30: Attributive adjektiver. Samspill mellom frekvens, antall og tekstlengde. De grå kurvene angir *antall* attributive adjektiver i teksten; teksten på den nederste, rette linjen har ingen attributive adjektiver, tekstene på kurven over har 1 attributivt adjektiv, tekstene på den neste har 2, etc.

Mange tekster har svært få attributive adjektiver; for eksempel har 22 tekster 4 eller færre. Lave antall medfører mer tilfeldig variasjon i tallene, og dette vil kunne maskere eventuelle systematiske tendenser. En sammenligning for effekt av kjønn og skriveverktøy der alle tekster med færre enn 5 attributive adjektiver er fjernet (figur 11-31), synes imidlertid ikke å tilsi at dette har noen vesentlig innvirkning på analysen.



Figur 11-31: Frekvens av attributive adjektiver. Interaksjon mellom skriveverktøy og kjønn. Til venstre alle tekstene, til høyre bare tekster med minst 5 attributive adjektiver.

Et annet usikkerhetsmoment er hva som faktisk er det validitetsmessig riktigste konstruktet for attributive adjektiver. Jeg var tidligere i dette delkapitlet inne på om det faktisk er antall per løpeord, eller om det heller kan være antall per substantiv eller antall per klausus. Korrelasjonen mellom håndtekstene og tastetekstene er temmelig lik i hvert tilfelle, $R \approx 0{,}39$ (se 11.3.3.2), $R \approx 0{,}40$, $R \approx 0{,}41$. Dette skulle ikke tilsi at noen av alternativene er overlegne noen av de andre med hensyn til validitet. Det er også svært sterk korrelasjon mellom de tre variantene av variabelen (se tabell 11-9 nedenfor), og dette skulle tilsi at det ikke spiller noen stor rolle hvilken variant som brukes.

Tabell 11-9: Attributive adjektiver med ulike målestokker. Pearsons korrelasjonskoeffisienter.

	Per løpeord	**Per substantiv**	**Per klausus**
Per løpeord	–		
Per substantiv	0,93	–	
Per klausus	0,98	0,90	–

Om man imidlertid legger en av de andre variabelvariantene til grunn for anova-analyse, blir ikke resultatet nødvendigvis det samme. Antall attributive adjektiver per substantiv gir en minimal adekvat modell med forskjell i tekstlengde som signifikant prediktor, $F \approx 5{,}87$, $p < 0{,}05$, der elever som skriver mye lengre med tastatur, har høyere frekvens av attributive adjektiver i tastetekstene (se appendiks A3). Antall attributive adjektiver per klausus gir derimot en nullmodell uten signifikans ($t \approx 1{,}28$, $p \approx 0{,}21$), altså det samme som med løpeord som målestokk. Korrelasjonen mellom disse to variabelvariantene er da også den sterkeste, $R \approx 0{,}98$.

Valg av målestokk kan altså påvirke resultatet av analysene, slik vi også så i 11.1.3.5 om korte subklaususer. Dette gir en viss usikkerhet i vurderingen av resultatet, men jeg mener det er gode grunner til å beholde antall løpeord som målestokk også for denne variabelen.

Frekvensen av attributive adjektiver synes intuitivt å være ganske lavt, selv om jeg ikke har sammenlignet med andre tekster. Tabell 11-10 nedenfor viser adjektivene som er mest brukt attributivt i korpuset:

Tabell 11-10: De mest frekvente attributive adjektiver i korpuset, lemmaformer

Leksem	Antall
stor	94
negativ	66
god	65
hel	57
gammal	41
ung	41
liten	34
forskjellig	32
litt	32
ny	30
sist	21
lang	17
første	14
tidlig	14
viktig	13
viss	13
dårlig	12
sterk	12
lett	10

Det påfallende frekvente \negativ\ er i stor grad hentet rett fra oppgavetekstens "negativ retning", i mange tilfeller omskrevet til \negativ utvikling\, men opptrer også i kollokasjoner som \negativ holdning\ og \negative konsekvenser\. \Hel\ opptrer 15 ganger som \helt feil\, mens mange av de andre forekomstene dreier seg om \hele tiden\ eller \hele dagen\ (totalt 20), altså ikke som prototypiske adjektiver. \Litt\ dreier seg i stor grad om en determinativ bruk. Det reelle antallet faktiske, funksjonelle adjektiver er dermed enda en del lavere enn det som kommer fram av det automatiske korpussøket, selv når jeg har fjernet \mye\, mange\, masse\, \få\, \faktisk\, \lite\ og ordenstall.

En potensiell årsak til lav frekvens kan være den korte tidsrammen, der mer mulighet for redigering av teksten kanskje kunne gitt en rikere tekst. En annen mulig årsak er rett og slett at tematikken og denne type argumenterende tekst ikke gjør det naturlig med mange attributive adjektiver.

11.4 Oppsummering av kapitlet

Jeg har i dette kapitlet analysert 8 ulike syntaktiske variabler, som alle er knyttet til syntaktisk kompleksitet, og der 7 av dem på ulike måter utfyller den ganske generelle variabelen gjennomsnittlig t-enhetslengde. Valget av de 7 utfyllende variablene er gjort med tanke på at de skal representere klaususlengde, antall ledd og leddlengde, men utvalget av

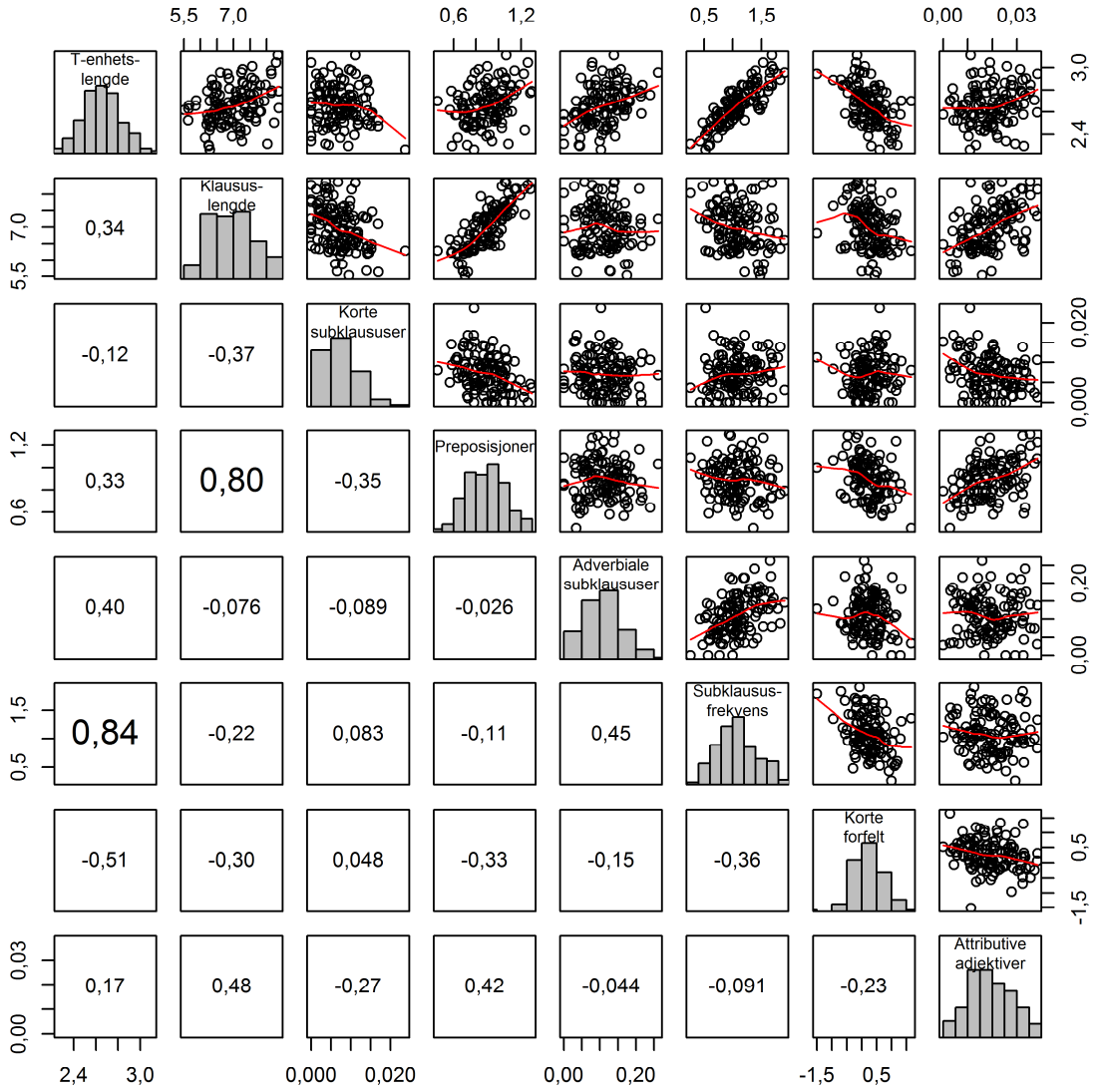
variabler har også vært avhengig av hva som er teknisk og praktisk mulig å trekke ut av korpuset på en automatisk eller semiautomatisk måte.

Tabell 11-11 nedenfor gir en oversikt over resultatene av variansanalysene for de 8 variablene.

Tabell 11-11: Resultater av variansanalysene av de syntaktiske variablene.

Variabel	**Resultat i tastetekster**	**Konklusjon**
T-enhetslengde	Lengre	Tvetydig
Klaususlengde	Ingen effekt	-
Korte subklaususer	Kortere tekster: Flere korte subklaususer Lengre tekster: Ingen effekt Gutter med liten tekstlengdeforskjell: Flere korte subklaususer (Gutter med stor tekstlengdeforskjell: Færre korte subklaususer)	Spontant - Spontant Redigert
Preposisjonsfraser	Kortere tekster: Færre preposisjonsfraser Lengre tekster: Flere preposisjonsfraser	Spontant Redigert
Adverbiale subklaususer	Stor tekstlengdeforskjell: Flere adverbiale subklaususer Liten tekstlengdeforskjell: Ingen effekt	Spontant -
Subklaususfrekvens	Flere i tastetekster	Spontant
Ettords forfelt	Jenter med liten tekstlengdeforskjell: Flere korte forfelt Resten: Færre korte forfelt	Spontant Redigert
Attributive adjektiver	Ingen effekt	-

De 7 utfyllende variablene er antatt å være partielle forklaringer på lengre t-enhetslengde, og man skulle på grunnlag av det forvente positiv korrelasjon mellom hver av dem og t-enhetslengde (log-transformert).



Figur 11-32: Korrelasjoner mellom de syntaktiske variablene. Pearsons korrelasjonskoeffisienter på 120 verdier som er parvis ikke-uavhengige.

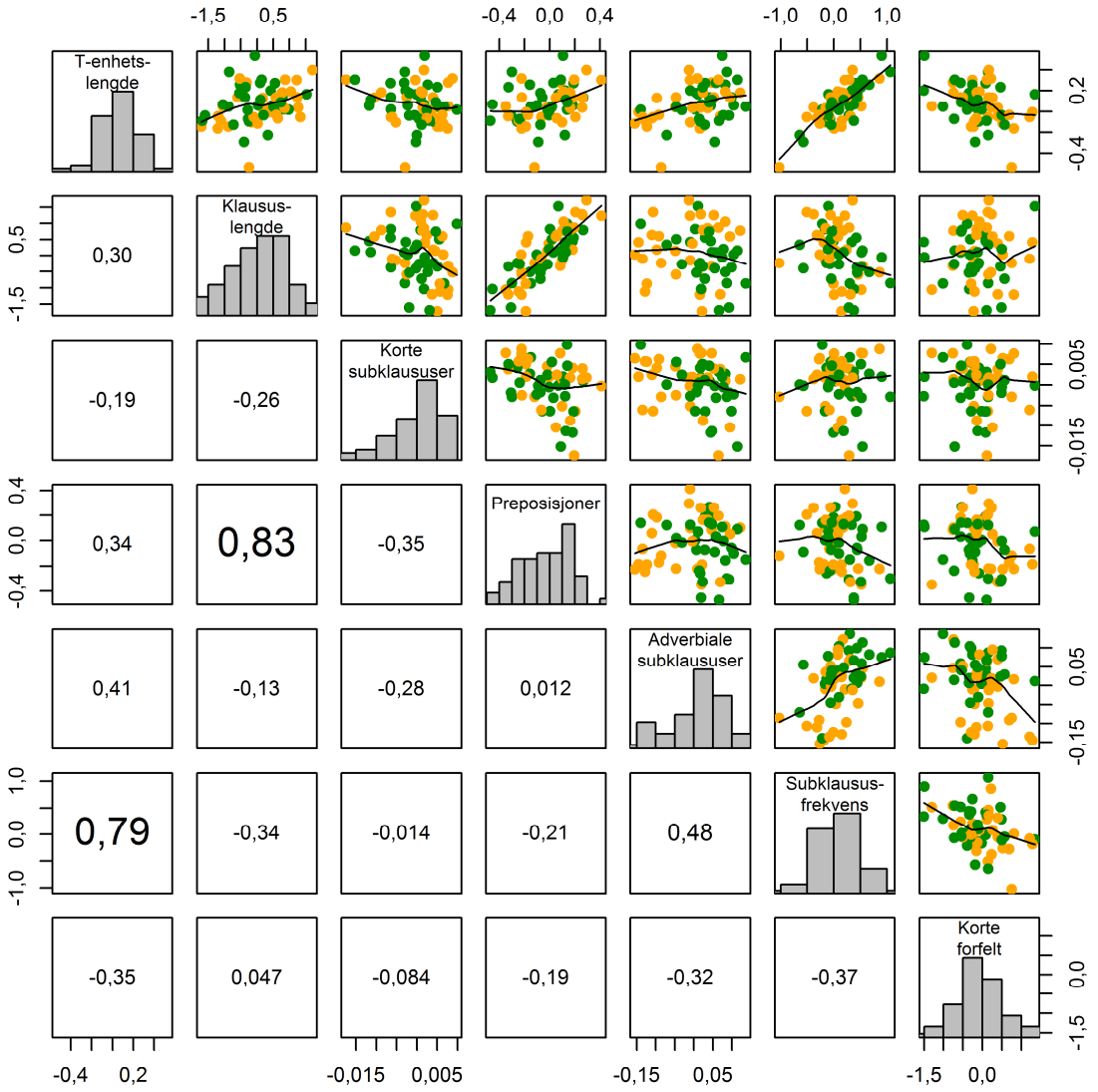
Krysskorrelasjonsdiagrammet i figur 11-32 viser korrelasjonene med t-enhetslengde og mellom de enkelte variablene. Diagrammet viser at 5 av variablene har signifikant positiv korrelasjon med t-enhetslengde. To av variablene, nemlig korte subklaususer og korte forfelt, er konseptuelt "snudd", slik at negative korrelasjonskoeffisienter faktisk gjenspeiler den egenskapen vi forventer. Attributive adjektiver og korte subklaususer har bare svak korrelasjon med t-enhetslengde; den er ikke-signifikant, men i utvalgene er retningen av korrelasjonen som forutsatt. At subklaususfrekvens er den egenskap som bidrar sterkest til t-enhetslengde stemmer godt med Hunts (1965) funn og er ikke direkte overraskende, mens det er litt vanskeligere å forklare at andelen korte forfelt skulle påvirke i så stor grad som $|R| \approx 0.48$. En mer plausibel forklaring er nok at det ikke er noen sterk kausalitet fra korte forfelt til t-enhetslengde, men at skrivemønsteret som delvis fanges opp gjennom denne

variabelen, henger sammen med flere andre trekk som også påvirker t-enhetslengden. Kanskje er det sterk sammenheng mellom korte forfelt og leddlengde generelt.

Også når vi ser på variabelen klaususlengde, går korrelasjonene i den forventede retning. Negativ korrelasjon med de to variablene som er knyttet til frekvens av subklaususer, er forventet ettersom underordnede klaususer ikke er regnet med i klaususlengden og dermed i en viss forstand bidrar negativt til lengden av den overordnede klaususen, slik det er forklart i 11.1.2.2. Påfallende sterk er korrelasjonen med preposisjonsfrekvens; når man ser bort fra subordinerte klaususer, er preposisjonsfraser det desidert viktigste bidraget til klaususlengde.

Blant de 6 gjenværende variablene er det generelt middels svak eller svakere korrelasjon, med unntak av korrelasjonen mellom de to målene som gjelder subklaususfrekvens. De lave korrelasjonskoeffisientene signaliserer at variablene representerer relativt uavhengige trekk i materialet, og de er derfor et signal om at variablene utgjør et fornuftig knippe av tekstlige egenskaper.

Hvis vi ser på tilsvarende krysskorrelasjonsdiagram for *differansene* mellom tastatur- og håndtekster i figur 11-33, kommer faktisk mange av de samme sammenhengene fram:



Figur 11-33: Korrelasjoner mellom differanseverdiene av de syntaktiske variablene. Pearsons korrelasjonskoeffisienter.

12 Prinsipalkomponentanalyse

I de foregående kapitlene har jeg studert variabler én og én for å avdekke hvordan de påvirkes av skriveverktøyet. Jeg har brukt visuelle teknikker og statistiske analyser med det formål å forstå de enkelte variablene, innimellom også med blikk på samspill mellom par av variabler. Som metode har jeg i stor grad benyttet forskjellige typer av hypotesetester, særlig anova-modellering med trinnvis modellreduksjon.

Som hypotesetesting er denne praksisen kritikkverdig, i og med at den medfører et relativt stort antall av hypotesetester, noe som fordrer justering av signifikansnivå (eller p-verdier) for å motvirke den økte risikoen for type-1-feil som forårsakes av gjentatt testing på det samme materialet (FWER). Med 13 hypotesetester for de 5 viktigste leksikalske variablene og de 8 syntaktiske variablene burde alfa-nivået reduseres til $0.05 * 0.05 / (1 – (1 – 0.05)^{13}) \approx 0.0051$. Som eksplorerende teknikk mener jeg den likevel kan forsvares, i og med at dette er et felt som så langt er lite utforsket, og at resultatene for de enkeltvariablene som er presentert her, kan danne grunnlag for hypoteser som kan testes mer metodologisk stringent på et annet materiale.

I dette kapitlet vil jeg imidlertid se på samspillet mellom variablene og helheten i materialet mer systematisk. Til dette bruker jeg en multivariat metode som kalles prinsipalkomponentanalyse (PCA), som jeg kombinerer med anova-analyse av samme type som i de foregående kapitlene. (Prinsippene for anova er forklart i 7.3.) Kapitlet begynner med en ikke-matematisk forklaring av hvordan PCA fungerer (12.1), før variablene som inngår i analysen, blir presentert (12.2). Selve analysen og resultatene blir gjennomgått i 12.3, før en avsluttende diskusjon (12.4).

12.1 Metode

Forklaringen av prinsippene bak PCA bygger på Baayen (2008, s. 118-).

Det ligger utenfor denne avhandlingen å forklare matematikken bak prinsipalkomponentanalyse, men prinsippet dreier seg om å redusere kaoset i et *n*-dimensjonalt rom dannet av *n* variabler ved å samle mest mulig variasjon i færrest mulig dimensjoner. Ved å definere nye dimensjoner som består av kombinasjoner av vektinger av de *n* variablene, søker man uavhengige dimensjoner der dimensjon 1 forklarer så mye av variasjonen i materialet som mulig, dimensjon 2 forklarer så mye av den gjenværende variasjonen som mulig, etc. Resultatet er et nytt *n*-dimensjonalt rom der mye av variasjonen forklares av de første få dimensjonene, mens den *n*-te dimensjonen forklarer svært lite av variasjonen. I dette nye rommet inngår hver av de opprinnelige variablene i hver av de nye dimensjonene, men med ulik vekting. Man kan også tenke på dimensjonene i det nye rommet som fremstått ved at dimensjonene i det opprinnelige rommet blir rotert for å finne den orienteringen som favner mest variasjon. I denne prosessen vil korrelasjon og mangel på korrelasjon mellom variablene spille en rolle, slik at variabler som korrelerer, gjerne vil ha stor vekt i samme dimensjon, mens variabler uten korrelasjon vil dominere ulike dimensjoner.

I PCA inngår altså ingen kunnskap om de prediktorvariablene som man har hypoteser om, og den skiller seg derfor fra klassifiseringsteknikker som diskriminantanalyse (Baayen, 2008, s. 154), som tar utgangspunkt i prediktorvariablene og hvordan de potensielt grupperer datapunktene. PCA skiller seg også fra faktoranalyse (Baayen, 2008, s. 126) ved at prinsippene for rotasjonen er forskjellige. Siktemålet ved rotasjonen i PCA er, som forklart over, å favne mest mulig variasjon i færrest mulig dimensjoner. Siktemålet ved rotasjonen i faktoranalyse er derimot å konstruere nye dimensjoner der de opprinnelige variablene har stor vekt bare i noen få dimensjoner, for slik å gjøre det enklere å fortolke innholdet i de enkelte dimensjonene. Det finnes ulike algoritmer for rotering i faktoranalyse, og de kan til dels gi ganske ulike resultater. Usikkerheten rundt dette har gjort at jeg har valgt å holde meg til PCA, som kan sees som teoretisk enklere.

Etter prinsipalkomponentanalysen har jeg så utført anova-modellering på de 4 viktigste av de resulterende dimensjonene, eller de "prinsipale komponentene". Disse 4 anova-analysene fungerer dermed i mye større grad som uavhengige hypotesetester, og deres p-verdier kan brukes uten justering av signifikansnivå. Man kunne nok med en viss rett hevde at også dette er repetert testing på det samme materialet, men ettersom de 4 dimensjonene har fullstendig fravær av indre korrelasjon, mener jeg det kan forsvares å se på dem som 4 uavhengige hypoteser med 4 uavhengige tester. Blant de resultatene jeg presenterer nedenfor, ville det motsatte standpunkt, altså FWER-korrigering, hatt konsekvens bare for 1 av de 3 signifikante funnene, nemlig for prediktoren kjønn i dimensjon 2, som har $p \approx 0.031$.

12.2 Variablene

Selv om PCA benytter en rent matematisk tilnærming til variasjon, så har det selvfølgelig konsekvenser for resultatet av de etterfølgende anova-analysene nøyaktig hvilke variabler man putter inn i den. I kapitlene om leksikalske variabler løfter jeg fram 5 variabler som jeg mener har best validitet, og som hver bidrar til et sammensatt bilde av de leksikalske egenskapene i tekstene, selv om det også er en del korrelasjon mellom noen av dem. De 5 variablene er gjennomsnittlig ordlengde (9.1), gjennomsnittlig ordlengde for leksikalske ord (9.3.2), leksikalsk tetthet (9.2), $\text{MOSTTR\_LL}_{W=50}$ (10.4.5) og korrigert $\text{log-TTR}_{1,3}$ (10.3.5). Det er derfor naturlig å la disse 5 variablene bidra i en PCA av materialet. I kapitlet om syntaktiske variabler fokuserer jeg på totalt 8 variabler, hvorav to – gjennomsnittlig t-enhetslengde (11.1.1) og gjennomsnittlig klaususlengde (11.1.2) – er en form for samlevariabler eller symptomvariabler for de andre, og som til dels korrelerer sterkt med enkelte av de syntaktiske enkeltvariablene. Jeg har derfor valgt å utelate disse to variablene fra prinsipalkomponentanalysen, og for å få en balanse mellom antall leksikalske og syntaktiske variabler har jeg dessuten latt være å inkludere den av de gjenværende 6 syntaktiske variablene som eksperimentet ikke lyktes å avdekke noen påvirkning på av skriveverktøyet (attributive adjektiver, 11.3.3). Dermed står jeg igjen med følgende 5 syntaktiske variabler for PC-analysen: frekvens av korte subklaususer (11.1.3), frekvens av preposisjoner (11.2.1), andel adverbiale subklaususer (11.2.2), antall subklaususer per t-enhet (11.3.1), og andel t-enheter med korte forfelt (11.3.2).

Tabell 12-1 nedenfor gir en oversikt over de 10 variablene og de forkortede variabelnavnene som er brukt i diagrammer og oversikter på de påfølgende sider.

Tabell 12-1: Oversikt over variablene i PC-analysen, med forkortelser

Variabel	**Forkortelse**
Gjennomsnittlig ordlengde	ordlengde
Gjennomsnittlig ordlengde for leksikalske ord	leksordlengde
Leksikalsk tetthet	leksord
MOSTTR_LL$_{W=50}$ (Lokal TTR)	MOSTTR
Korrigert log-TTR (Global TTR)	logTTR
Frekvens av korte subklaususer	korte_subkl
Frekvens av preposisjoner	prep
Andel adverbiale subklaususer (per subklausus)	adv
Antall subklaususer per t-enhet	subkl
Andel t-enheter med korte forfelt	korte_forfelt

I PC-analysen er det de aktuelle differanseverdiene mellom håndtekster og tastetekster for hver variabel som inngår. Det vil si at jeg har brukt de samme differanseverdiene som i anova-analysene i de foregående kapitlene, inkludert logit-baserte verdier der dette er aktuelt. Dessuten er alle verdier sentrert og normalisert til *z*-verdier, altså slik at hver variabel har middelverdi = 0 og lik varians (standardavvik = 1). Normaliseringen er gjort for å unngå at ulik målestokk skal medføre at visse variabler får uforholdsmessig stor vekt i analysen.

Krysskorrelasjonstabellen i tabell 12-2 nedenfor viser samspillet mellom de valgte 10 variablene. Mange av korrelasjonene under har vært presentert i tidligere kapitler, men i denne tabellen kommer dessuten sammenhengen mellom de leksikalske variablene og de syntaktiske variablene fram, i de 5x5 fargelagte rutene nederst til venstre.

Tabell 12-2: Korrelasjoner mellom leksikalske variabler og syntaktiske variabler. Verdiene er Pearsons korrelasjonskoeffisienter.

Pearsons R	Ordlengde	Leksikalsk ordlengde	Leksikalsk tetthet	MOSTTR	logTTR	Korte subklaususer	Preposisjoner	Adverbiale subklaususer	Subklausus-frekvens	Korte forfelt
Ordlengde	–									
Leksikalsk ordlengde	0,85	–								
Leksikalsk tetthet	0,68	0,33	–							
MOSTTR	0,34	0,53	0,04	–						
logTTR	0,54	0,57	0,29	0,85	–					
Korte subklaususer	–0,12	–0,17	–0,18	–0,26	–0,25	–				
Preposisjoner	0,41	0,34	0,40	0,32	0,43	–0,35	–			
Adverbiale subklaususer	–0,15	0,01	–0,34	0,19	0,12	–0,09	–0,03	–		
Subklaususfrekvens	–0,15	–0,01	–0,37	0,04	–0,05	0,08	–0,11	0,45	–	
Korte forfelt	–0,22	–0,29	–0,03	–0,18	–0,19	0,06	–0,33	–0,14	–0,36	–

Det går fram at det ikke finnes noen sterke korrelasjoner mellom de leksikalske variablene og de syntaktiske variablene. Den største korrelasjonskoeffisienten er 0,43, mellom preposisjoner per klausus og log-TTR. Faktisk er mange av de sterkeste korrelasjonene nettopp mellom preposisjonsvariabelen og de leksikalske variablene, noe som viser at bruk av preposisjoner henger sammen med leksikalsk variasjon, tetthet og spesifisitet. At for eksempel gjennomsnittlig ordlengde henger sammen med preposisjonsbruk, viser at det neppe er egenskapene ved preposisjonene direkte dette gjelder, men egenskapene til utfyllingene i preposisjonsfrasene.

Ellers er de sterkeste tendensene den negative korrelasjonen mellom leksikalsk tetthet og de to subklaususvariablene. Dette støtter Hallidays tese om at informasjonell tetthet og syntaktisk kompleksitet er motpoler blant stilistiske trekk.

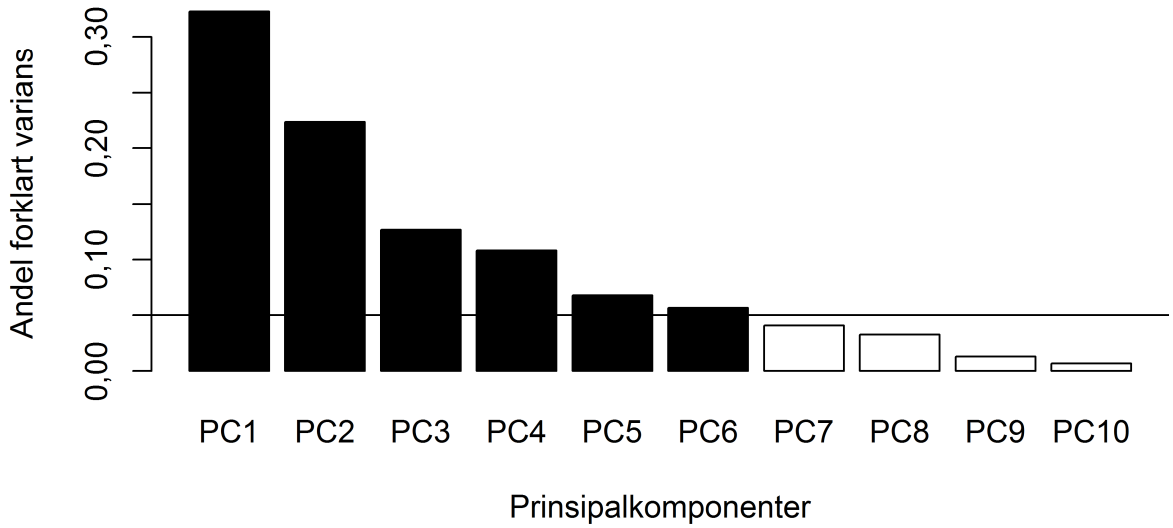
Tabell 12-3 nedenfor viser korrelasjonen mellom de leksikalske variablene og de to syntaktiske samlevariablene, gjennomsnittlig t-enhetslengde og gjennomsnittlig klaususlengde. Det er bare lave verdier for korrelasjonskoeffisientene som gjelder t-enhetslengde, mens det er verdt å legge merke til at alle de leksikalske variablene korrelerer positivt med klaususlengde. Det tyder på at gjennomsnittlig klaususlengde faktisk *er* et modenhets- eller ferdighetsmål, mens gjennomsnittlig t-enhetslengde kan være uttrykk for ganske ulike egenskaper. Korrelasjonskoeffisientene for klaususlengde er også påfallende like i størrelse, mellom 0,32 og 0,40.

Tabell 12-3: Korrelasjon mellom de to syntaktiske samlevariablene og de leksikalske variablene. Verdiene er Pearsons korrelasjonskoeffisienter.

Pearsons R	**Ordlengde**	**Leksikalsk ordlengde**	**Leksikalsk tetthet**	**MOSTTR**	**logTTR**
T-enhetslengde	0,07	0,19	– 0,16	0,22	0,17
Klaususlengde	0,40	0,34	0,37	0,32	0,37

12.3 Analyse og resultater

Prinsipalkomponentanalysen resulterer i et system av 10 dimensjoner der 6 dimensjoner er nødvendig før man kommer til en dimensjon som forklarer mindre enn 5% av den totale variasjonen, som illustrert i figur 12-1 nedenfor.



Figur 12-1: Forklart varians i prinsipalkomponentene

Skreddiagrammet gir en oversikt over dimensjonene, sortert i rekkefølge etter andel variasjon de forklarer, mens tabell 12-4 nedenfor gir de nøyaktige tallene for andel forklart variasjon samt akkumulert forklart variasjon.

Tabell 12-4: Forklart og akkumulert forklart varians i prinsipalkomponentene

Dimensjon	Forklart var	Akkumulert var
PC1	0,322	0,322
PC2	0,224	0,546
PC3	0,126	0,673
PC4	0,108	0,781
PC5	0,068	0,849
PC6	0,057	0,905
PC7	0,041	0,946
PC8	0,033	0,979
PC9	0,013	0,993
PC10	0,007	1

Det går fram av tabellen at de to første dimensjonene er de klart viktigste og til sammen forklarer 55% av variasjonen. De to neste dimensjonene er ganske like i viktighet, med mellom 10% og 13% hver, mens dimensjon 5 og 6 ligger nær grensen på 5%, som gjerne brukes som kritisk nedre grense for analyse av PCA-dimensjoner (Baayen, 2008, s. 121). Det er utfordrende å forklare eller tolke et system av mange dimensjoner, og jeg kommer i fortsettelsen til å konsentrere meg om de to viktigste dimensjonene, men også kort omtale dimensjon 3 og 4. Krysskorrelasjonstabellen med Pearsons korrelasjonskoeffisienter i tabell 12-5 nedenfor viser at disse 4 dimensjonene er fullstendig uavhengige, slik forutsetningen for PCA er.

Tabell 12-5: Krysskorrelasjoner som viser prinsipalkomponentenes uavhengighet. Verdiene er Pearsons korrelasjonskoeffisienter.

Pearsons R	**PC1**	**PC2**	**PC3**	**PC4**
PC1				
PC2	0,00			
PC3	0,00	0,00		
PC4	0,00	0,00	0,00	

Tabell 12-6 nedenfor viser vektingen av de 10 variablene i hver av de fire viktigste dimensjonene. Fortegnet viser retningen den enkelte variabelen påvirker den enkelte dimensjon, mens absoluttverdien av tallet viser vekten. I analysen har jeg ikke gjort noe forsøk på å vende variabler etter den retningen hypotesen har; i stedet lar jeg PCA avgjøre hvordan dimensjonen er påvirket av den enkelte variabel, slik at jeg i etterkant kan tolke variabelens egenskaper.

Tabell 12-6: Vekting av de enkelte variablene i de 4 første prinsipalkomponentene

	PC1	**PC2**	**PC3**	**PC4**
korte_subkl	0,174	0,161	0,676	0,133
prep	– 0,066	– 0,368	– 0,460	0,394
adv	– 0,458	0,075	– 0,049	– 0,005
subkl	– 0,331	0,345	0,154	0,203
korte_forfelt	0,232	0,011	– 0,207	– 0,797
ordlengde	– 0,003	– 0,579	0,369	– 0,074
leksordlengde	– 0,341	– 0,296	0,342	– 0,286
leksord	0,257	– 0,476	0,022	0,175
MOSTTR	– 0,485	– 0,016	– 0,090	– 0,172
logTTR	– 0,418	– 0,252	0,006	– 0,063

For å få oversikt over de enkelte dimensjonene er det nyttig å sortere variablene etter vektingens absoluttverdien, slik det går fram av tabellene i tabell 12-7 nedenfor. Når man studerer enkeltvariablenes påvirkning på dimensjonen, er det fornuftig å sette en nedre grense for variabelens vekt. Det finnes imidlertid ingen prinsipiell verdi for akkurat hvor lite vekt en variabel skal ha før den kan sies å være uinteressant, og jeg har markert to alternative grenseverdier 0,3 og 0,2 i tabellene under, ved hjelp av ulik fargelegging.

Tabell 12-7: De 4 første prinsipalkomponentene med sorterte variabler

	PC1		**PC2**		**PC3**		**PC4**
MOSTTR	– 0,485	ordlengde	– 0,579	korte_subkl	0,676	korte_forfelt	– 0,797
adv	– 0,458	leksord	– 0,476	prep	– 0,460	prep	0,394
logTTR	– 0,418	prep	– 0,368	ordlengde	0,369	leksordlengde	– 0,286
leksordlengde	– 0,341	subkl	0,345	leksordlengde	0,342	subkl	0,203
subkl	– 0,331	leksordlengde	– 0,296	korte_forfelt	– 0,207	leksord	0,175
leksord	0,257	logTTR	– 0,252	subkl	0,154	MOSTTR	– 0,172
korte_forfelt	0,232	korte_subkl	0,161	MOSTTR	– 0,090	korte_subkl	0,133
korte_subkl	0,174	adv	0,075	adv	– 0,049	ordlengde	– 0,074
prep	– 0,066	MOSTTR	– 0,016	leksord	0,022	logTTR	– 0,063
ordlengde	– 0,003	korte_forfelt	0,011	logTTR	0,006	adv	– 0,005

Tabellene viser at det er nødvendig med 4 dimensjoner for at alle variablene skal være med i minst 1 dimensjon med vekt > 0,3. Dersom man velger kritisk grense på 0,2, er det tilstrekkelig med 3 dimensjoner. Det går også fram at de 2 første dimensjonene trekker på 8 av variablene med vekt > 0,3, henholdsvis 9 med vekt > 0,2. De to variablene som ikke får særlig vekt i de to første variablene, får svært mye vekt i hver sin av de påfølgende dimensjoner, henholdsvis 0,68 (korte subklaususer) og –0,80 (korte forfelt). At nettopp disse to variablene havner i dimensjoner for seg selv, henger sammen med at det er disse to som har de svakeste korrelasjonene med de andre variablene, slik det går fram av krysskorrelasjonstabellen i tabell 12-2 ovenfor. Det er også verdt å merke seg at de leksikalske variablene og de syntaktiske variablene ikke grupperer seg i ulike dimensjoner, men at de blander seg i alle de viktigste dimensjonene. Dette underbygger det inntrykk at jeg har kommet fram til et sett av variabler som bidrar med hver sine fasetter av et komplekst bilde.

For hver av disse dimensjonene regner PCA ut nye verdier for hver elev, basert på vektingen av hver enkelt variabel i hver dimensjon (se i appendiks A6). På disse verdiene har jeg utført anova-analyser med trinnvis modellreduksjon på samme måte som for enkeltvariablene i de foregående kapitlene (forklart i 7.3.1 og 7.3.2), altså med de 4 dikotome prediktorvariablene kjønn, ferdighet, total tekstlengde og forskjell i tekstlengde, med antall prediktorinteraksjoner begrenset til 2:

```
lm(pc1~(kjønn+ferdighet+lengde+forskjell)^2)
lm(pc2~(kjønn+ferdighet+lengde+forskjell)^2)
lm(pc3~(kjønn+ferdighet+lengde+forskjell)^2)
lm(pc4~(kjønn+ferdighet+lengde+forskjell)^2)
```

Trinnvis modellreduksjon etter de prinsipper som er forklart i 7.3.1 ovenfor, fører til følgende minimale adekvate modeller:

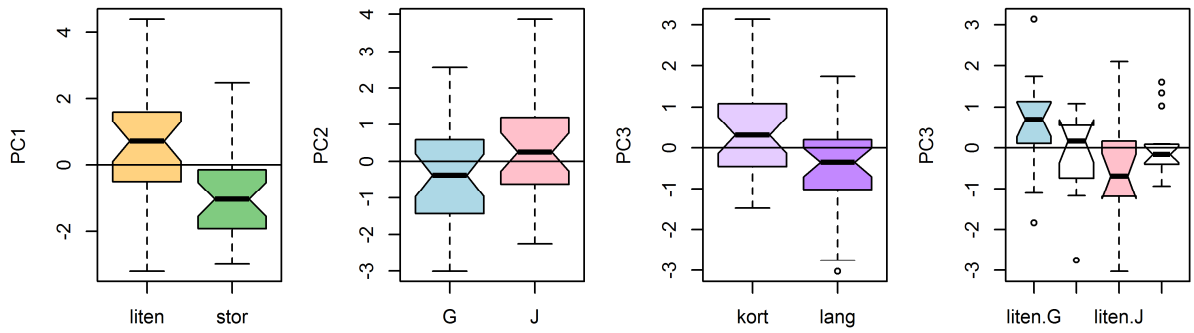
```
PC1
            Df Sum Sq Mean Sq F value   Pr(>F)
forskjell    1  34.18   34.18   12.71 0.000735 ***
Residuals   58 155.91    2.69
---
Residual standard error: 1.64 on 58 degrees of freedom
Multiple R-squared:  0.1798,    Adjusted R-squared:  0.1657

PC2
            Df Sum Sq Mean Sq F value Pr(>F)
kjønn        1  10.32  10.324   4.908 0.0307 *
Residuals   58 122.00   2.103
---
Residual standard error: 1.45 on 58 degrees of freedom
Multiple R-squared:  0.07802,   Adjusted R-squared:  0.06212

PC3
               Df Sum Sq Mean Sq F value  Pr(>F)
kjønn           1   3.00   2.997   2.887 0.09494 .
lengde          1   8.30   8.302   7.998 0.00652 **
```

```
forskjell         1   0.04   0.044   0.042 0.83809
kjønn:forskjell   1   6.14   6.136   5.911 0.01833 *
Residuals        55  57.09   1.038
---
Residual standard error: 1.019 on 55 degrees of freedom
Multiple R-squared:  0.2344,    Adjusted R-squared:  0.1787
F-statistic:  4.21 on 4 and 55 DF,  p-value: 0.004808
```

For PC1 er forskjell i tekstlengde signifikant ($F \approx 12{,}71$, $p < 0{,}001$); for PC2 er kjønn signifikant ($F \approx 4{,}91$, $p \approx 0{,}031$); for PC3 er tekstlengde ($F \approx 8{,}00$, $p \approx 0{,}0065$) og interaksjonen mellom kjønn og forskjell i tekstlengde ($F \approx 5{,}91$, $p \approx 0{,}018$) signifikant, mens omnibus-verdiene for den reduserte modellen for PC3 er $F \approx 4{,}21$, $p \approx 0{,}0048$. `Gvlma` (se 7.2.2.4) rapporterer at premissene for anova er oppfylt i alle tre tilfeller. (Se appendiks A4.) For interaksjonen i dimensjon 3 rapporterer Tukeys HSD-test at kun forskjellen mellom jenter som har liten tekstlengdeforskjell, og gutter som har liten tekstlengdeforskjell, er signifikant, $p \approx 0.026$ (se appendiks A5). For dimensjon 4 blir resultatet en ikke-signifikant nullmodell.[46]



Figur 12-2: Signifikante prediktoreffekter i de tre første prinsipalkomponentene. I diagrammet til høyre illustrerer de fargelagte boksene den signifikante kjønnsforskjellen blant elever med stor tekstlengdeforskjell.

Boksdiagrammene i figur 12-2 ovenfor viser de signifikante effektene. I diagram nummer 1 fra venstre ser vi at elever med liten tekstlengdeforskjell har høyere PC1-verdier i tastetekstene, mens elever med stor tekstlengdeforskjell har lavere PC1-verdier i tastetekstene. I diagram nummer 2 ser vi at jenter har litt høyere PC2-verdier i tastetekster, mens gutter har lavere PC2-verdier i tastetekstene, men effekten for gutter eller jenter er ikke særlig stor; det er *forskjellen* mellom kjønnene som er signifikant. I diagram nummer 3 ser vi at elever som skriver generelt kort, har høyere PC3-verdier i tastetekstene enn elever som skriver generelt langt. I diagram 4 ser vi at blant elever med liten tekstlengdeforskjell er det forskjell mellom gutter og jenter ved at gutter har høyere PC3-verdier i tastetekstene, mens

[46] At nullmodellen er ikke-signifikant, følger som en nødvendighet av at differansevariablene er sentrert før PC-analysen. En ettutvalgs t-test, som er ekvivalent med nullmodellen, gir $t \approx 0$ for *alle* prinsipalkomponentene.

jenter har lavere PC3-verdier i tastetekstene. Vi merker oss ellers at ferdighet er den eneste prediktoren som ikke spiller en rolle i de 4 første dimensjonene i PC-analysen.

Tabell 12-8 nedenfor gir oversikt over p-verdier og effektmål (Cohens *d*) for hver prinsipalkomponent. Effektene er mellom 0,5 og 1 standardavvik, størst for effekten av tekstlengdeforskjell i PC1, mens effekten er svakest for kjønnsforskjellen i PC2, der p-verdien også nærmer seg det valgte signifikansnivået på 0,05. For PC1 er imidlertid p-verdien klart under ethvert konvensjonelt valgt signifikansnivå, og det er hevet over enhver rimelig tvil at det finnes en reell forskjell mellom elever som skriver mye lengre på tastatur, og de andre når det gjelder hvordan de utnytter skriveverktøyene. Dessuten er forskjellene klare og tydelige for begge de signifikante prediktorene i PC3.

Tabell 12-8: Signifikans og effekt i de første prinsipalkomponentene

PC	**Prediktor**	**p**	**Cohens *d***	**Høyere verdi**
1	Tekstlengdeforskjell	< 0,001	0,94	Liten tekstlengdeforskjell
2	Kjønn	0,031	– 0,58	Jenter
3	Lengde	0,007	0,74	Korte tekster
3	Kjønn (liten tekstlengdeforskjell)	0,018	0,86	Gutter
4	–	–	–	–

Vi kan da oppsummere effektene på denne måten:

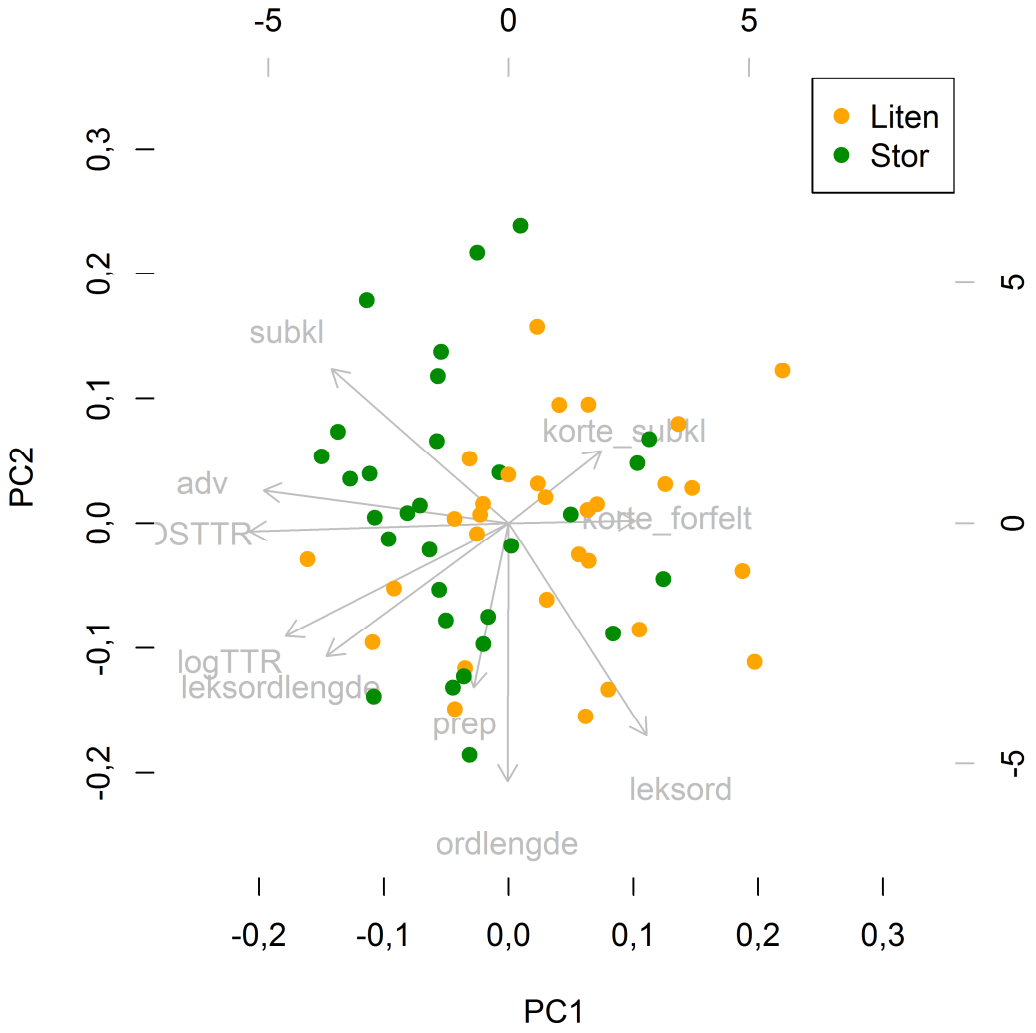
- PC1: Elever som skriver mye lengre på tastatur, har i forhold til dem som ikke skriver mye lengre på tastatur, i sine tastetekster i forhold til i sine håndtekster:
  - mer leksikalsk variasjon
  - større andel av adverbiale subklaususer
  - lengre leksikalske ord
  - høyere frekvens av subklaususer
  - (lavere leksikalsk tetthet)
  - (lavere andel av korte forfelt)
- PC2: Gutter har i forhold til jenter i sine tastetekster i forhold til i sine håndtekster
  - lengre ord
  - høyere leksikalsk tetthet
  - flere preposisjoner per klausus
  - lavere frekvens av subklaususer
  - (lengre leksikalske ord)
  - (mer global leksikalsk variasjon)
- PC3: Elever som skriver generelt lengre, har i forhold til dem som skriver generelt kortere, i sine tastetekster i forhold til i sine håndtekster
  - Lavere andel av korte subklaususer
  - Flere preposisjoner per klausus
  - kortere ord
  - kortere leksikalske ord
  - (høyere andel korte forfelt)

- PC3: Det samme gjelder jenter med liten forskjell i tekstlengde i forhold til gutter med liten forskjell i tekstlengde.

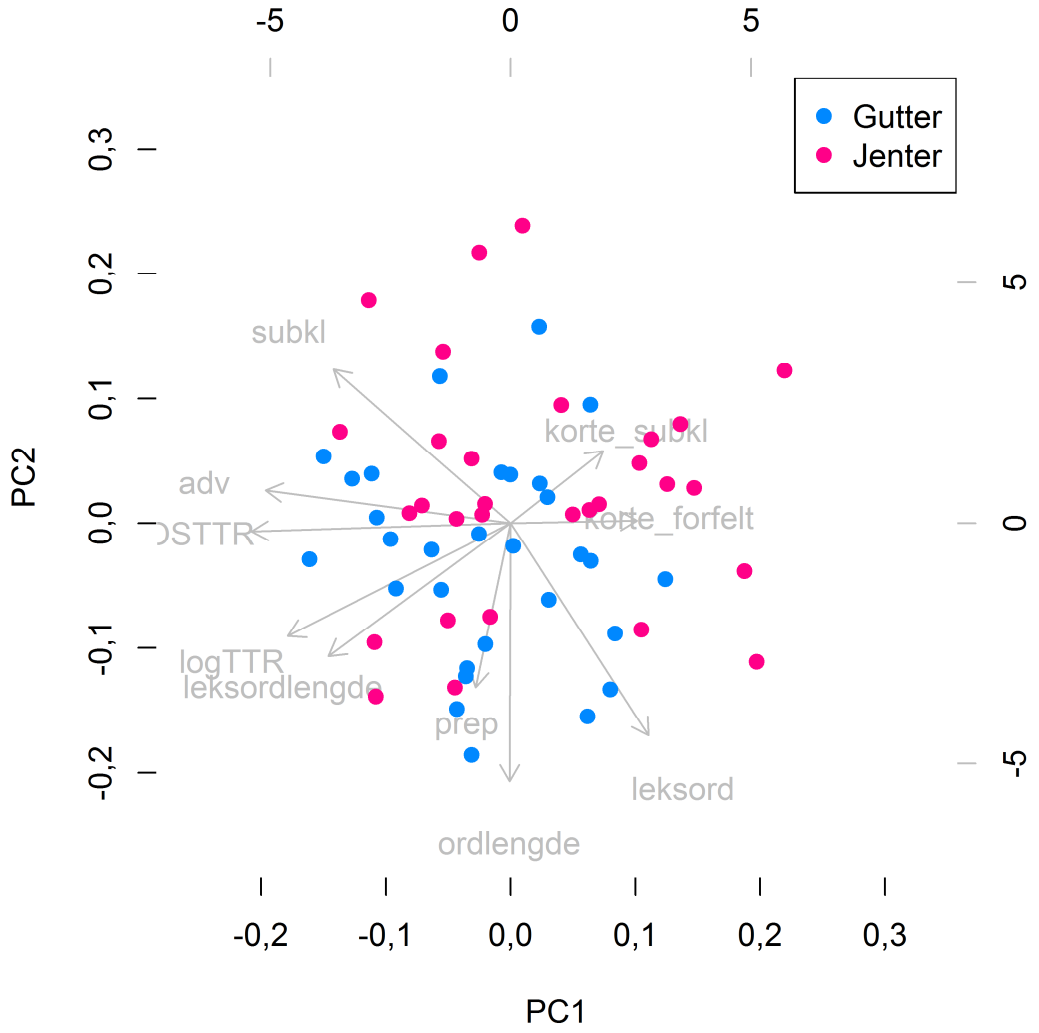
Det er viktig nå å huske på at tendensene bare gjenspeiler *forskjeller* mellom utvalgene. PC-analysen kan ikke alene fortelle oss for eksempel at gutter har lengre ord i sine tastetekster enn i sine håndtekster; den kan bare si at forskjellen er større enn hos jentene. Kanskje er det slik at guttene har forskjell og jentene *ikke*, kanskje er det slik at guttene har positiv forskjell og jentene negativ, eller kanskje er det slik at guttene *ikke* har forskjell, mens jentene har negativ forskjell. Dette kan ikke PC-analysen fortelle oss, og vi må gå tilbake til analysene av de enkelte variablene i foregående kapitler for å avdekke effekten for de enkelte variablene.

I oversikten ovenfor kan det være vanskelig å forstå sammenhengene mellom de variablene som utgjør de viktigste variablene i PC3, og ettersom den forklarte variansen for denne dimensjonen også er såpass lav (13%), fokuserer jeg heretter på de 2 første dimensjonene, som altså til sammen forklarer omtrent 55% av variasjonen i differanseverdiene.

I de to spredningsdiagrammene i figur 12-3 og figur 12-4 nedenfor er de to dimensjonene plottet mot hverandre, med PC1 langs den horisontale aksen og PC2 langs den vertikale aksen. I diagrammene er også variablene tegnet inn som vektorer, der vektorens lengdekomponent i henholdsvis x- og y-retning indikerer variabelens vekt i PC1 og PC2. I det første diagrammet er forskjell i tekstlengde indikert ved hjelp av fargemarkering av punktene, mens i det andre diagrammet er kjønn markert på tilsvarende måte.



Figur 12-3: Spredningsdiagram for PC1 og PC2, fordelt etter tekstlengdeforskjell



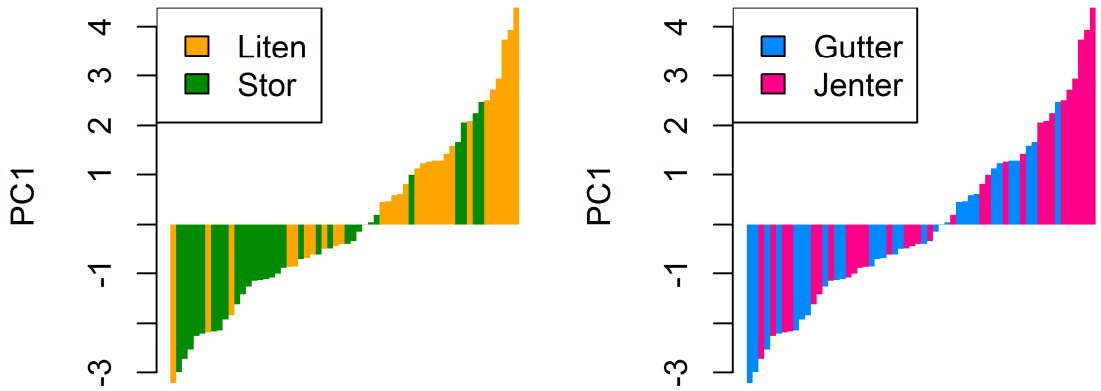
Figur 12-4: Spredningsdiagram for PC1 og PC2, fordelt etter kjønn

Det første diagrammet viser tydelig en forskjell mellom elever som skriver mye lengre på tastatur, og de andre. De grønne punktene er forskjøvet mot venstre i diagrammet, altså i retning av blant annet mer leksikalsk variasjon og større andel adverbiale subklaususer i tastetekstene, mens de gule punktene har en overvekt i den høyre delen av diagrammet, altså i retning av flere korte forfelt og flere korte subklaususer i tastetekstene. Men som det går fram av tabell 12-7 ovenfor, har disse to variablene lite vekt i PC1, så viktigere er det at de gule punktene befinner seg i et område som er negativt definert i forhold til de vektorene som peker mot venstre, altså blant annet *mindre* leksikalsk variasjon og *lavere* andel adverbiale subklaususer i tastetekstene.

I det andre diagrammet er ikke tendensen like lett å få øye på, men det er en overvekt av gutter i det nedre del av diagrammet, og en overvekt av jenter i den øvre del av diagrammet. Det illustrerer altså at guttene har blant annet forholdsvis lengre ord og flere preposisjoner

per klausus i tastetekstene, mens jentene har forholdsvis kortere ord og færre preposisjoner per klausus i tastetekstene.

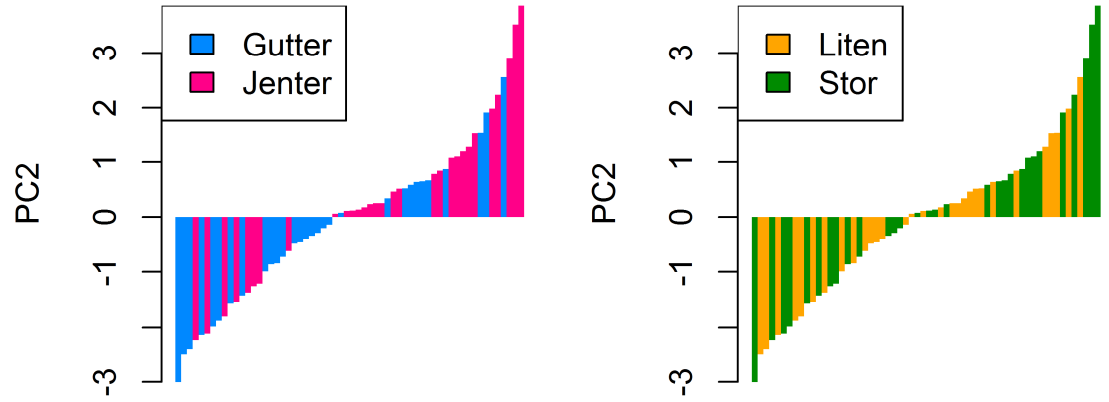
Hvis vi ser på skreddiagrammer for bare PC1 i figur 12-5 nedenfor, skilt på tekstlengdeforskjell i diagrammet til venstre og på kjønn i diagrammet til høyre, ser vi tydelig effekten av tekstlengdeforskjell:



Figur 12-5: Sorterte verdier for PC1 fordelt etter tekstlengdeforskjell (til venstre) og kjønn (til høyre)

Blant de elevene som har det klareste utslaget for PC1, er det bare 7 elever som bryter mønsteret, 3 med liten tekstlengdeforskjell og 4 med stor tekstlengdeforskjell. I diagrammet til høyre er det liten synlig kjønnseffekt blant de negative verdiene, men det er en klar overvekt av jenter i den positive enden. Det vil altså si at også PC-analysen finner den tendensen vi har sett gjennom flere av enkeltvariablene, nemlig at det er en håndfull jenter (pluss 1 gutt i dette tilfellet), omtrent en tredjedel, som ser ut til å skille seg ut fra resten av utvalget. Det er imidlertid for mye blanding i resten av verdiområdet til at en anova-analyse avdekker denne tendensen, og kjønnsprediktoren framstår ikke som signifikant i PC1.

Om vi ser på tilsvarende diagrammer for PC2, i figur 12-6 nedenfor, er det som ventet ingen synlig effekt av tekstlengdeforskjell, hverken i de negative eller de positive verdiene. Kjønnseffekten er imidlertid ganske klar, særlig blant de største positive verdiene, der jentene dominerer. Av en eller annen grunn ser det også ut til at kjønnene grupperer seg hver for seg rundt nullpunktet, men det er liten grunn til å anta at dette er noe annet enn en tilfeldig effekt.



Figur 12-6: Sorterte verdier for PC2 fordelt etter kjønn (til venstre) og tekstlengdeforskjell (til høyre)

12.4 Diskusjon

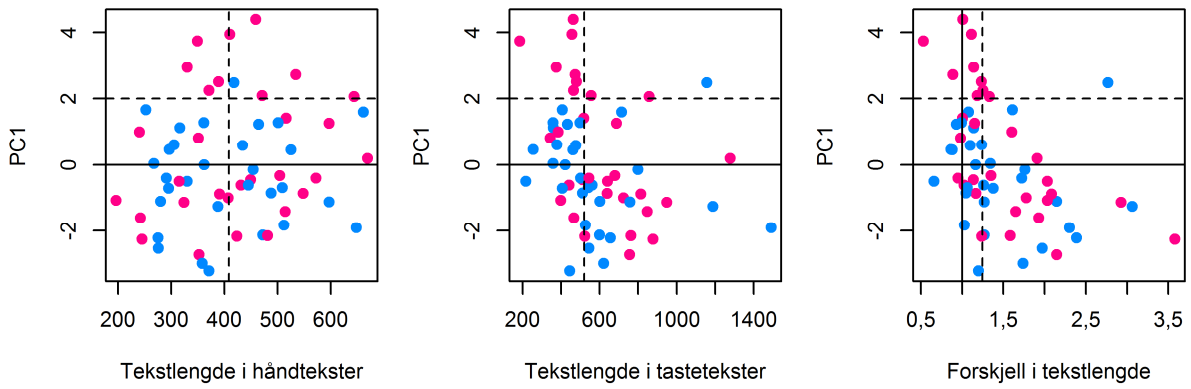
PC1 og PC2 er altså matematisk uavhengige, men når vi ser på tolkningen av variablene som inngår i dem, er det vanskelig å skille mellom dem og tillegge dem ulike fortolkninger eller funksjoner, slik Biber (1988, s. 104-) gjør i sin faktoranalyse. Det er jo også slik at to variabler inngår (negativt) i begge dimensjonene, mens to andre inngår med ulikt fortegn. Kanskje kunne vi hevde at (negativ) PC1 representerer rikdom og kreativitet og flyt, med leksikalsk variasjon, leddsetninger og lengre leksikalske ord, mens (negativ) PC2 representerer mer redigert tekst, med lengre ord, høyere informasjonstetthet, flere preposisjonsfraser og mindre kompleks klaususstruktur. Med en slik tolkning kan det virke som at de som skriver lengre på tastatur, skriver "rikere", kanskje fordi de produserer raskere, slik jeg trekker fram i hypotesen i 5.1, eller kanskje delvis også fordi disse elevene faktisk behersker verktøyet bedre, og dermed får tid til å redigere inn mer leksikalsk variasjon, for eksempel. Dessuten kan det virke som at gutter utnytter verktøyet til å redigere i større grad enn jenter, og at kanskje særlig en del jenter produserer mer spontane språktrekk på tastatur.

Jeg hadde i utgangspunktet ingen konkret hypotese om forskjeller mellom kjønnene relatert til effekt av skriveverktøy, og jeg ble litt overrasket over dette resultatet. Man kan spekulere over om det skyldes påvirkning fra andre skrivesituasjoner, og at flere jenter enn gutter på den tiden dataene ble samlet inn, brukte pc-en til sosiale medier. Fra innholdet i elevenes tekster om "Bøker eller data" ser det ut til at elevene oppfatter at det er en slik forskjell i bruksområdene, der guttene spiller mer, mens jentene bruker sosiale medier mer. Men i den grad det er mulig å trekke informasjon om bruk av sosiale medier ut av spørreskjemaene, så er det ingen kjønnsforskjeller i hvordan gutter og jenter selv rapporterer dette. Blant de 10 elevene med høyest verdier på dimensjon 2, 7 jenter og 3 gutter, er det heller ikke slik at de

skiller seg systematisk fra resten av elevgruppen, hverken i bruksfrekvens, bruksmengde eller bruksområder.[47]

Gutter og jenter er selvfølgelig også elever som skriver mye lengre eller ikke så mye lengre på tastatur, og i figur 12-5 ovenfor går det klart fram at en god del av elevene som ikke øker tekstlengden så mye på tastatur, og som *ikke* produserer rikere på tastatur, er jenter. 9 av de 10 elevene med høyest PC1-verdier er jenter, og i det høyre spredningsdiagrammet i figur 12-7 nedenfor går det fram at ingen av disse jentene har mer enn moderat økning i tekstlengde i tastetekstene – riktignok ligger ikke alle under medianen. Den enslige gutten med høy PC1-verdi har en påfallende annerledes posisjon langt til høyre, ettersom han skriver nesten 3 ganger så langt på tastatur som for hånd. I de to andre diagrammene i figuren er PC1 relatert til tekstlengde i henholdsvis håndtekster (til venstre) og tastetekster (i midten). Figuren illustrerer i hvor sterk grad forskjell i tekstlengde sammenfaller med lengde i tastetekster ($R \approx 0{,}68$), mens det er liten grad av slikt sammenfall i mønster med lengde i håndtekster. Jentene med høye PC1-verdier har også i stor grad korte tastetekster, selv om det ikke gjelder absolutt alle.

Vi skal være forsiktige med å legge kvalitetsverdier i dimensjonene; vi har for eksempel sett i 10.3.5 at leksikalsk variasjon ikke nødvendigvis henger sammen med tekstkvalitet. Men vi ser altså at jenter dominerer både det vi har kalt den "fattige" enden av PC1 og det vi har kalt den "spontane" enden av PC2, men at det hovedsakelig er *forskjellige* jenter som dominerer i de to dimensjonene. Totalt er 14 av de 18 elevene som enten har høyest PC1-verdier eller høyest PC2-verdier, jenter; 2 jenter ligger i begge områdene. Ut fra analysen av de leksikosyntaktiske trekkene kan det dermed se ut til at det spesielt er en del av jentene som ikke utnytter skriveverktøyet i samme grad som resten av elevgruppen.



Figur 12-7: Sammenheng mellom tekstlengdevariablene, kjønn og PC1

Til slutt vil jeg minne om at PCA ikke benytter kunnskap om hvordan eventuelle prediktorvariabler som kjønn klassifiserer eller grupperer de språklige variablene. Målet for

[47] Disse 10 elevene er elevene med følgende elev-ID, sortert i stigende rekkefølge etter verdiene i dimensjon 2: 285 202 294 312 302 305 248 298 205 243.

en PCA er kun å lage en modell som best mulig sorterer variasjonen mellom de *språklige* variablene. Jeg tror at kjønnstendensen i materialet kunne ha kommet enda klarere fram dersom valg av rotasjon hadde vært motivert av å maksimere forklaringskraften til prediktorene. Om vi ser en gang til på spredningsdiagrammet i figur 12-4 ovenfor, er det ikke vanskelig å ane et mønster der guttene befolker den nedre, venstre delen av diagrammet, mens jentene er i overtall i den øvre, høyre delen. Jeg tror at en rotasjon av koordinatsystemet ca. 45 grader mot høyre bedre ville ha fått fram kjønnsforskjellen, i stedet for å fordele kjønnseffekten på begge aksene slik PC-analysen gjør. En naiv "rotasjon" i form av en summering av verdiene for PC1 og PC2 øker Cohens *d* for kjønn fra henholdsvis 0,45 og 0,58 til 0,74, noe som bekrefter det visuelle inntrykket diagrammet gir.

Hvis vi vender tilbake til PC3, så er det altså slik at elever som generelt skriver kortere tekster, har høyere andel av korte klaususer, lavere frekvens av preposisjonsfraser og lengre ord og leksikalske ord i tastetekstene i forhold til elever som generelt skriver lengre tekster. Med det motsatte perspektivet kan det se ut til at elever som skriver langt, i større grad fyller klaususene med mer innhold på tastatur, men at innholdet ikke er så spesifikt. Det er ganske sterk korrelasjon mellom PC3 og klaususlengde ($R \approx -0{,}53$), men tendensen er langt svakere når vi sammenligner de to elevsegmentene med hensyn til klaususlengde, så forklaringen er mer nyansert enn som så. Jeg har tidligere pekt på at det er en viss sammenheng mellom tekstlengde og ferdighet; blant de som skriver lengst, er 19 sterke og 11 middels elever. Ferdighet fremstår imidlertid ikke som en signifikant faktor i modellen og skiller langt svakere på PC3 enn tekstlengde (henholdsvis $t \approx 1{,}68$ og $t \approx 2{,}81$). Det kan tenkes at tekstlengde er et godt symptom på skriveferdighet, slik det går fram av analysene fra KAL-prosjektet og andre prosjekter (8.4.1), men det er uansett vanskelig å forklare akkurat den effekten som framkommer i PC3, som dessuten representerer bare 13% av variasjonen.

Den eneste faktoren som ikke framstår som signifikant i noen av de 4 første prinsipalkomponentene, er skriveferdighet. Ferdighet spiller også liten rolle i de fleste enkeltvariablene, så med de enkelte variabelanalysene som utgangspunkt er ikke dette så overraskende. At ferdighet ikke spiller noen rolle for hvordan elevene utnytter skriveverktøyet, er imidlertid i strid med hva jeg hadde ventet meg. Jeg regnet med at sterkere elever har en fordel når det gjelder å utnytte mer avanserte hjelpemidler, og at skriveproduktene ville bære spor av dette i sine leksikosyntaktiske trekk. Det kan tenkes at mangelen på effekt av ferdighet skriver seg fra at de sterkere elevene først og fremst henter ut en fordel i redigering på de øvre tekstnivåer, og at den type redigering ikke gjenspeiles i de leksikosyntaktiske trekkene. Kanskje er ikke forskjeller i redigering på mikronivå avhengig av ferdighet i særlig grad, men i større grad knyttet til andre personlige egenskaper, for eksempel egenskaper som kommer til uttrykk som kjønnsforskjeller i analysene. Vi skal heller ikke glemme to vesentlige egenskaper ved variabelen ferdighet slik den er målt i denne undersøkelsen. For det første er den selvrapportert, noe som trolig har resultert i en viss unøyaktighet. For det andre er det slik at det er ganske lite spredning i variabelen. 53 av 60 elever har enten 4 eller 5 som norskkarakter, og denne spredningen er kanskje rett og slett for liten til å få fram noen effekt på de aktuelle variablene.

13 Oppsummering og diskusjon

Hovedproblemstillingen i denne avhandlingen er hvorvidt og eventuelt hvordan leksikosyntaktiske trekk i elevtekster påvirkes av skriveverktøy, altså av om de er skrevet på tastatur eller for hånd, og hvordan en slik påvirkning eventuelt arter seg ulikt for ulike segmenter av elever. I innledningen skisserer jeg fem ulike nytteverdier av en slik undersøkelse:

1. Økt kunnskap om hva slags faktorer som påvirker språkproduksjon generelt.
2. Økt kunnskap om hvordan digitale omgivelser påvirker språkproduksjon spesielt.
3. Bedre kunnskapsbasert grunnlag for å drive skriveopplæring med ulike skriveverktøy.
4. Økt kunnskap om leksikosyntaktiske trekk i elevers skriving generelt.
5. Økt kunnskap om metoder for kvantitative korpusundersøkelser av språktrekk.

Disse punktene griper inn i hverandre på ulike måter og er ikke alltid lett å holde fra hverandre. Når det gjelder punktet som dreier seg om økt kunnskap om leksikosyntaktiske trekk i elevtekster generelt, viser jeg til de enkelte resultatene som er gjengitt i analysekapitlene over; jeg kommer ikke til å gjengi dem her.

I dette kapitlet vil jeg først samlet oppsummere resultatene av analysene som skulle besvare hovedproblemstillingen, og som er knyttet til alle de tre øverste punktene i listen over. Deretter vil jeg peke på noen av gevinstene som er knyttet til metodeutvikling og dessuten diskutere noen begrensninger ved de metodene som er brukt. Til slutt vil jeg kort skissere hva jeg mener er fornuftige neste skritt på vegen i dette feltet.

13.1 Oppsummering

Jeg har gjennom analysene og drøftingene i denne avhandlingen endt opp med 13 leksikosyntaktiske variabler som jeg anser som de mest verdifulle av dem jeg har vurdert. Det er 5 leksikalske variabler og 8 syntaktiske variabler:

- gjennomsnittlig ordlengde
- gjennomsnittlig ordlengde i leksikalske ord
- leksikalsk tetthet
- global TTR (justert for tekstlengde)
- lokal TTR (tekstlengde nøytralisert)

- gjennomsnittlig t-enhetslengde
- gjennomsnittlig klaususlengde
- frekvens av korte subklaususer (per løpeord)
- antall preposisjonsfraser per klausus
- antall adverbiale subklaususer per klausus

- antall subklaususer per t-enhet
- andel t-enheter med kort forfelt
- frekvens av attributive adjektiver (per løpeord)

For analyseresultatene for de enkelte variablene viser jeg til presentasjonen i foregående kapitler.

13.1.1 Resultater

Hypotesen for effekten av skriveverktøyet var knyttet til at verktøyet kan påvirke språklige trekk i en planlagt, redigert, "skriftlig" retning, eller i en spontan, "muntlig" retning. Ut fra denne tankegangen kan det se ut som at enkeltresultatene kan samles i fem kategorier.

For det første er det noen variabler som ikke ser ut til å bli påvirket av skriveverktøyet; i hvert fall har ikke analysene avdekket noen slik påvirkning. Dette gjelder gjennomsnittlig klaususlengde og frekvens av attributive adjektiver. Det kan være flere årsaker til en slik mangel på funn. For det første kan det tenkes at ulike deler av skrivesituasjonen påvirker variabelen i ulike retninger for ulike elever, og at disse effektene utligner hverandre. For eksempel har jeg flere ganger vært inne på at forskjellen mellom de to skriveoppgavene og forskjeller i tekstlengde kan maskere eventuelle resultater. For det andre har vi sett at valg av målestokk for et trekk kan ha avgjørende innvirkning på analyseresultatet, og kanskje har jeg valgt en målestokk som i liten grad fanger opp det fenomenet jeg ønsket å analysere. For det tredje kan det hende at effekten er så liten at styrken i eksperimentet og analysene ikke er kraftig nok til å avdekke den, men vi kan selvfølgelig heller ikke se bort fra muligheten av at dette er en variabel som rett og slett ikke blir påvirket av skriveverktøyet.

For det andre er det noen variabler som blir påvirket i spontan retning av tastaturskriving. Dette var som forutsatt i den mest overgripende hypotesen, nemlig at økt skrivehastighet vil gi mer spontane trekk i teksten. Dette gjelder for subklaususfrekvens. Det er også slik at gjennomsnittlig t-enhetslengde er høyere i tastetekstene enn i håndtekstene, men det er mer problematisk å karakterisere dette som et ubetinget spontant trekk uten å se nærmere på hva slags språktrekk som bidrar til den økte t-enhetslengden. I utgangspunktet hadde jeg regnet med at flere variabler ville vise påvirkning i spontan retning av tasting, men ut fra resultatene som er oppsummert nedenfor, kan det virke som at skriveverktøyet – ikke overraskende ut fra tidligere drøfting – påvirker språktrekkene i ulike retninger, og at ulike elever og kanskje ulike teksttyper blir påvirket på ulike måter. Dermed blir det få entydige resultater.

For det tredje er det noen variabler som blir påvirket i spontan retning av tastaturskriving for jenter og i planlagt, redigert retning av tastaturskriving for gutter. Dette gjelder ordlengde, ordlengde i leksikalske ord, global TTR og korte forfelt. Jeg hadde i utgangspunktet ingen konkrete hypoteser knyttet til kjønn, men ettersom det er tydelige kjønnsforskjeller i norske skoleprestasjoner generelt og skriveferdigheter spesielt, var det naturlig å ha med denne faktoren i datainnsamlingen og analysene. Ettersom denne effekten gjelder såpass mange variabler, synes den kjønnsforskjellen som er avdekket, å ha et ganske solid fundament. Riktignok er det slik at det er sterk korrelasjon mellom de to ordlengdemålene, så de er ikke

to helt uavhengige variabler. Vi har også sett at man kanskje bør være litt forsiktig med å karakterisere høyere global TTR som et planlagt, redigert trekk. Hvis vi holder fast ved forskningslitteraturens konklusjoner om TTR, har vi likevel en ganske klar kjønnsforskjell der noen av effektene er blant de sterkeste i hele materialet.

Også for en del variabler der det ikke er noen kjønnsforskjell i påvirkning fra skriveverktøyet, er det slik at gutter skriver med mer planlagte, redigerte trekk generelt enn jentene. Dette gjelder klaususlengde, preposisjonsfraser, adverbiale klaususer og lokal TTR. Litt uformelt kan det se ut som om gutter putter mer materiale inn i segmentene sine, på flere ulike måter. Kanskje er det slik at gutter generelt har et litt annet tekstideal enn jenter, og at de i noen sammenhenger utnytter pc-verktøyet til å nærme seg dette idealet.

Men kjønnsforskjellen dreier seg ikke bare om at gutter skriver mer planlagt eller redigert med pc; det er også slik at jenter skriver mer spontant på pc, og dette er vanskeligere å forklare, synes jeg. Spørreskjemasvarene indikerer ingen klare kjønnsforskjeller når det gjelder hva de bruker pc til hjemme, utover en overvekt av gutter som spiller spill, men noen av elevtekstene er inne på at jentene bruker pc-en mer til sosiale medier enn gutter:

> Faktisk så tror jeg at det er en større del av jenter som sitter på dataen enn gutter i de siste årene, fordi jenter sitter for det meste på Facebook og MSN og den slags, mens gutter spiller på spillkonsoller og sånt. [A1-247]

> Gutter spiller i større grad PC-spill enn jenter. World of warcraft og andre krigsspill er det guttene som er storforbrukere av. Det er noen jenter som spiller også, men de fleste jenter driver i større grad med hjemmesider, blogging og chatbaserte internettsider som Facebook og Nettby. [A1-250]

Kanskje er det slik at jenter tar med seg et skrivemønster fra sosiale medier over i skoleskrivingen, og at denne overføringen i størst grad påvirker skrivingen på tastatur? Dette er imidlertid bare spekulering, og dette eventuelle årsaksforholdet må undersøkes med en annen forskningsdesign enn gjennom korpusstudier. Det er heller ikke slik at elevene er entydig enige om den virkelighetsbeskrivelsen som kommer fram i de to sitatene over:

> men i ungdommenes verden er det små forskjeller [mellom gutter og jenter]. Guttene lager seg sider på Facebook, Nettby og blogger, akkurat slik som jentene. Sitter på MSN, chatter og møter venner på nett. [A1-203]

> Halvparten eller over halvparten av de som spiller World of Warcraft er faktisk jenter. Både jenter og gutter spiller spill, surfer på nettet, skriver og gjør lekser på dataen og de leser bøker, på og av dataen. [A1-263]

For det fjerde er det noen variabler der påvirkningen fra skriveverktøyet har sammenheng med hvorvidt eleven har mye lengre tastetekst enn håndtekst eller ikke. Elever som skriver mye lengre på tastatur, har høyere global TTR, høyere lokal TTR og mer adverbiale subklaususer i tastetekstene, mens elever som skriver omtrent like langt med de to verktøyene, har den motsatte effekten. I utgangspunktet stemmer ikke resultatene for TTR-

målene med hypotesen; lengre tastetekster skulle tilsi raskere produksjon, som igjen skulle føre til mer spontane trekk i teksten og ikke mer planlagte, redigerte trekk. Kanskje er det imidlertid slik at de elevene som skriver mye lengre tastetekster enn håndtekster, har bedre ferdigheter med pc-verktøyet, og at de greier å utnytte disse ferdighetene både til å produsere raskere og til å planlegge og/eller redigere teksten. Slik blir kanskje effekten av redigeringen sterkere enn effekten av hastigheten. Det kan også være at de kortere tekstene er uttrykk for lav motivasjon eller "å ikke ha noe å skrive om". Lav motivasjon fører til mindre arbeid med tekstene, noe som kan resultere i mer spontane språktrekk. Og igjen må vi huske på at effekten av oppgaveteksten kan forstyrre eller maskere effekten av skriveverktøy. Imidlertid stemmer resultatet med hypotesen for adverbiale subklaususer; mer bruk av slike er et spontant trekk som vi forventer å finne i tastetekstene til de som skriver mye lengre på tastatur. At begge TTR-målene utviser de samme interaksjonsmønstrene med skriveverktøy som adverbiale subklaususer, forsterker usikkerheten rundt disse variasjonsmålene som planlagte, redigerte trekk. Temmelig beskjedne korrelasjonskoeffisienter mellom adverbiale subklaususer og henholdsvis global og lokal TTR ($R \approx 0{,}12$; $R \approx 0{,}19$) tyder imidlertid på at relasjonen mellom variablene er sammensatt, og de bekrefter de svakhetene ved TTR-baserte mål som jeg har vært inne på tidligere.

Den helhetlige tilnærmingen i prinsipalkomponentanalysen bekrefter kjønnsforskjellene i effekten av skriveverktøy som vi finner i enkeltvariablene. Prinsipalkomponentanalysen bekrefter også at effekten av skriveverktøy i sterk grad er påvirket av tekstlengdeforskjell, men denne påvirkningen involverer et noe overraskende knippe av variabler som gjør at det vil være nødvendig å gå nærmere inn på de enkelte variablene for å forstå de mekanismene som er i virksomhet. Måten jeg har brukt prinsipalkomponentanalysen på, gjør at den ikke er egnet til å bekrefte de to første delfunnene over, altså de som er knyttet til enten *ingen* effekt av skriveverktøy eller *allmenn* effekt av skriveverktøy.

For det femte finnes det for noen variabler interaksjonsmønstre som ikke passer inn i de fire kategoriene over. Dette gjelder frekvens av korte subklaususer, som har effekter av både total tekstlengde og interaksjon mellom tekstlengdeforskjell og kjønn, leksikalsk tetthet, som viser interaksjon mellom ferdighet og tekstlengdeforskjell, antall preposisjonsfraser per klausus, som er høyere i tastetekstene til de som skriver langt, og kortere i tastetekstene til de som skriver kort, og andel t-enheter med kort forfelt, som har en interaksjon mellom kjønn og tekstlengdeforskjell som for så vidt passer i både kategori 3 og kategori 4 over.

13.1.2 Metoder

De viktigste resultatene fra dette doktorgradsarbeidet er de kvantitative tekstegenskapene som jeg har oppsummert i forrige underkapittel. Men etter min mening bidrar også avhandlingen med viktig metodeutviklingsarbeid. Dette gjelder først og fremst metoder knyttet til leksikalske variabler, men også i en viss grad til syntaktiske variabler.

Jeg drøfter inngående to ulike strategier for å nøytralisere den tekstlengdeavhengigheten som er en inherent egenskap ved TTR-målet, som i ulike varianter er ganske utbredt som et mål på leksikalsk variasjon eller diversitet i tekster. Den ene strategien benytter transformasjoner

for å rette opp både helningsgrad og kurvatur i sammenhengen mellom TTR og tekstlengde. Den andre strategien benytter gjennomsnitt av kortere tekstsegmenter av samme lengde for å gjøre TTR-målet uavhengig av tekstlengde. Jeg drøfter og sammenligner ulike tilnærminger til begge strategier og ender opp med å velge én tilnærming til hver strategi som jeg rapporterer resultater fra, og som inngår i den helhetlige prinsipalkomponentanalysen. De to tilnærmingene korrelerer ganske sterkt og representerer dermed beslektede egenskaper i tekstene, men jeg argumenterer for at de representerer henholdsvis global og lokal variasjon, og at de dermed representerer to noe ulike leksikalske egenskaper, selv om de ikke er to helt uavhengige dimensjoner. Jeg valgte også å måle ordformvariasjon med den ene variabelen og lemmaformvariasjon blant bare de leksikalske ordene med den andre variabelen. Jeg mener at begge disse TTR-baserte målene er validitetsmessig og reliabilitetsmessig verdifulle variabler, men det er nok også slik at dette området ikke er ferdig utforsket, og det gjenstår en del arbeid.

Dessuten representerer TTR-baserte variabler bare en del av et komplekst bilde av hva leksikalsk variasjon er, og jeg utforsker andre tilnærminger til variasjon enn repetisjonsbaserte, blant annet en variabel basert på entropi. Den variabelen jeg kommer fram til, har ikke de nødvendige egenskapene, men jeg tror denne delen av analysen kan være et bidrag i et felt som bør utforskes videre. Jeg gjør også forsøk med frekvensbaserte variabler for leksikalsk spesifisitet som jeg tror kan videreutvikles til valide mål for relevante tekstegenskaper.

Når det gjelder den syntaktiske delen av arbeidet, tror jeg at særlig diskusjonen rundt målestokk for slike språkbrukstrekk er verdifull. I tidligere arbeider har vi sett at det velges målestokker kanskje basert først og fremst på praktiske argumenter, men uten at konsekvensene av valgene drøftes eller ulike varianter sammenlignes empirisk. Jeg tror at vi trenger en mer bevisst forståelse av hvilke følger slike valg kan ha for kvantitative resultater, og en mer bevisst diskusjon av forholdet mellom målestokken og forskningsspørsmålet.

Som grunnlag for de grammatiske analysene har jeg i utstrakt grad støttet meg på en kombinasjon av automatiske morfologiske annoteringer og manuelle syntaktiske annoteringer. Det er mitt inntrykk at dette var et gunstig valg. Denne kombinasjonen av automatisk og manuell annotering gir svært rike muligheter for å gjenfinne relevant informasjon på ulike nivåer i tekstene. Jeg tror også det er riktig å si at den tidkrevende manuelle annoteringen var nødvendig for å oppnå nødvendig presisjon i den syntaktiske segmenteringen, selv om nyere arbeid innen syntaktisk parsing og trebanking kanskje vil gjøre slikt manuelt arbeid overflødig.

13.2 Diskusjon

Enkle *power*-beregninger (Howell, 2007, s. 231-) viser at en parametrisk ettutvalgstest med N = 60, som i praksis er det vi benytter for å finne allmenne effekter i hele elevutvalget, har *power* = 0,8 for $d \approx 0{,}36$. Det vil si at dersom effekten i populasjonen er $d \approx 0{,}36$, har vi 80% sjanse for å avkrefte nullhypotesen med et utvalg på N = 60, eller en $\beta = 0{,}2$ risiko for type-

2-feil. Dette nivået for *power* er et vanlig valg og et fornuftig kompromiss mellom risiko for type-1-feil og risiko for type-2-feil, ifølge Howell. Dette vil altså innebære at vi har mindre sjanse for å bekrefte effekter som er mindre enn $d \approx 0{,}36$. Intuitivt og ut fra de variabelegenskapene vi har sett i de foregående eksemplene, virker dette som et fornuftig nivå; verdien ligger omtrent midt mellom det Cohen kalte en liten effekt (0,2) og en medium effekt (0,5), og effekter på mindre enn ca. et tredjedels standardavvik tror jeg ikke har særlig stor betydning i denne sammenhengen. Når vi derimot ser etter interaksjonseffekter mellom skriveverktøy og elevfaktorer som kjønn eller ferdighet, blir bildet noe annerledes. I dette tilfellet gir *power* = 0,8 og n = 30 i hvert utvalg $d \approx 0{,}72$. For interaksjoner mellom to elevfaktorer (n = 15) er tilsvarende verdi for $d \approx 1{,}02$.[48] Det vil si at i disse tilfellene blir risikoen for å bekrefte en falsk nullhypotese $\beta > 0{,}2$ dersom den reelle effekten i populasjonen er henholdsvis $d < 0{,}72$ og $d < 1{,}02$. Jeg vil si at effekter på for eksempel $d = 0{,}5$ er av interesse, og eksperimentet ser ut til å hva for lite styrke til å avdekke slike forskjeller med tilfredsstillende sikkerhet. I parentes bemerket kan man spørre seg om det er *d*-verdier basert på differanseverdiene eller effektstørrelser basert på standardavviket i variabelverdiene som er best egnet for å danne seg et bilde av styrken i tendensene. Kanskje ville det ha vært nyttig også å kikke på *d*-verdier basert på variabelverdiene.

Beslektet med spørsmålet om *power* og de følger det har for risikoen for type-2-feil, er risikoen for å utelate faktiske effekter i en trinnvis reduksjon av anova-modeller. Vi har gjennom analysene sett eksempler på at forskjellige variabler med sterke korrelasjoner seg imellom har resultert i ulike minimale modeller. Med såpass lav *power*, og altså såpass stor risiko for å bekrefte falske nullhypoteser, er det naturlig at en slik trinnvis reduksjonsprosess også vil forkaste blant annet interaksjonseffekter vi potensielt ville være interessert i. Til en viss grad kunne dette ha vært hjulpet på ved å kombinere modellreduksjon med modellkonstruksjon, men i det store og hele er det manglende *power* som ligger under disse problemene.

Slik jeg ser på denne undersøkelsen, har den både eksplorerende og hypotesetestende trekk. Jeg bruker signifikansberegninger knyttet til statistiske modeller som gjennomgående metode, men det teoretiske og empiriske grunnlaget for de hypotesene jeg tester, er ikke veldig solid. Det er dermed naturlig å se på resultatene også delvis i et eksplorerende perspektiv, og jeg bruker dette som begrunnelse for å ikke korrigere signifikansnivået for *familywise error rate*. I et eksplorerende perspektiv er det imidlertid problematisk at testenes *power* er såpass svak.

I et hypotesetestende perspektiv er det viktig å være oppmerksom på at utvalget av elever er balansert og ikke representativt. Det innebærer at vi må være forsiktigere med å generalisere til populasjonen enn dersom utvalget hadde vært representativt. Graden av balansering i

---

[48] Disse grenseverdiene er framkommet ved oppslag i tabeller i Howell (2007). Av grunner som trolig har å gjøre med at Howell og jeg bruker noe avvikende definisjoner av Cohens *d*, stemmer ikke Howells grenseverdier nøyaktig med resultatene i denne avhandlingen.

forhold til et representativt utvalg fra skolens elever er imidlertid moderat, og jeg tror ikke det ligger noen stor risiko i å generalisere resultatene.

Jeg tror Halliday har rett i at kompleksitet ikke er et endimensjonalt fenomen, og jeg mener dette synet får klar støtte i tidligere empiriske studier, og at det også bekreftes av analysene i denne avhandlingen. Denne oppfatningen har vært med å danne grunnlaget for utvalg av de leksikosyntaktiske variablene som er blitt analysert i prosjektet. Den har vært en sterk motivasjon for å inkludere både leksikalske og syntaktiske variabler i modellene; jeg tror at leksikalske og syntaktiske variabler danner ulike og gjerne uavhengige dimensjoner i et tekstlig rom, men også at visse leksikalske og syntaktiske variabler er gjensidig avhengige av hverandre, påvirker hverandre og henger sammen. Jeg mener derfor at en språktrekksanalyse av et tekstmateriale basert på bare leksikalske eller bare syntaktiske variabler ville danne kun et halvt bilde av landskapet, og at det er nødvendig for å få fram helheten å se på kombinasjonen av slike variabler. Termen *leksikosyntaktisk* er også valgt nettopp for å fremheve at variablene henger sammen, og at de er *mer* enn leksikalske og syntaktiske variabler hver for seg. Når det gjelder det konkrete valget av de fem leksikalske og de åtte syntaktiske variablene, har tidligere studier, særlig Bibers, dannet et viktig grunnlag, men valgene er også gjort på litt ulik måte. Når det gjelder de leksikalske variablene, er en vesentlig del av avhandlingen dedikert til utvikling, sammenligning og evaluering av ulike variabler. En del forsøk har vist seg å være mindre fruktbare av ulike årsaker, men jeg har dessuten forsøkt å konstruere et sett av leksikalske variabler som til sammen representerer ulike dimensjoner av kompleksitet eller stil, altså variabler som i liten grad korrelerer med hverandre. Når det gjelder de syntaktiske variablene, har formålet med utvalget vært det samme, men disse variablene er i større grad valgt ut fra et tankeeksperiment om av hva slags syntaktiske kompleksitetsdimensjoner som kan tenkes å eksistere, heller enn en utvikling, sammenligning og evaluering av variablene. I noen tilfeller har jeg imidlertid sammenlignet og evaluert varianter av variabler. Til sammen var tanken at de 13 variablene skulle danne et 13-dimensjonalt rom uten for sterke korrelasjoner seg imellom, og altså at de åtte syntaktiske variablene skulle komplettere bildet som var påbegynt av de fem leksikalske variablene. Utvalget av 13 variabler kan åpenbart diskuteres. Når det gjelder den leksikalske delen av bildet, tror jeg det er blitt åpenbart gjennom drøftingene i denne avhandlingen at det må finnes viktige leksikalske egenskaper som ikke representeres av noen av de leksikalske variablene i undersøkelsen. Når det gjelder den syntaktiske delen av bildet, er det tilsvarende klart at det ikke ligger like mye arbeid bak utviklingen av disse variablene; utvalget er til en viss grad tilfeldig. Det er viktig å understreke at andre leksikosyntaktiske variabler kunne ha vært valgt, og at et slikt valg sannsynligvis ville ha gitt andre analyseresultater. Hvor annerledes resultatene ville ha blitt, vet vi ikke.

Beslektet med diskusjonen i forrige avsnitt er de diskusjonene jeg har gjennomført undervegs omkring hva slags målestokk som er mest valid for de ulike fenomenene. Dette er ikke spørsmål som har absolutte svar, og de er dessuten avhengige av det forskningsspørsmål og den hypotese som er valgt. Et eksempel er valget av MOSTTR over MATTR, der argumenter av ulike typer finnes for hver av de to variantene, og valget kanskje

like gjerne kunne ha falt på MATTR. Også valg av t-enhet og finitt klausus som analyseenheter har konsekvenser; å inkludere for eksempel infinitivkonstruksjoner og eventuelt andre infinitte verbfraser i klaususbegrepet, slik Halliday gjør, vil resultere i andre variabelverdier og potensielt andre analyseresultater.

Et gjennomgående prinsipp i operasjonaliseringen av begreper og forskningsspørsmål har imidlertid vært det pragmatiske. Jeg har søkt å operasjonalisere på en slik måte at jeg har kunnet utføre mest mulig omfattende analyser basert på minst mulig manuelt annoterings- eller analysearbeid. Derfor har jeg i så stor grad som mulig støttet meg på den automatiske analysen fra taggerprogrammet, og heller vurdert hva slags analysegrep som synes å gi minst feilmargin, enn å gå inn å gjøre omfattende manuelle korrigeringer i taggingen. Et eksempel er at jeg utelukker alle adverb fra mengden av leksikalske ord heller enn å vurdere hver enkelt lemmaform. Jeg har også lagt vekt på å dokumentere hvordan korpussøkene er gjennomført, og hvordan jeg har manipulert trefflistene. Dette har dessuten verdi med tanke på etterprøvbarhet og mulighet for å sammenligne resultater mellom ulike studier.

Til slutt vil jeg nevne at jeg nok har mer kunnskap om elevene i undersøkelsen enn jeg har greid å utnytte. Å bruke hva elevene faktisk skriver om i tekstene, ville kunne supplere bildet av dem, men jeg tenker først og fremst på de spørreskjemadataene jeg har samlet inn, og som jeg nok i større grad kunne ha benyttet, om ikke i selve de statistiske analysene, så i hvert fall i fortolkning av resultatene.

13.3 Videre arbeid

Statistiske språktrekksanalyser av norske elevtekster er ikke den forskningen vi har sett mest av de siste årene. Jeg vil gjerne se replikering av det eksperimentet jeg har beskrevet i denne avhandlingen, også med andre språkvariabler, andre teksttyper, andre elevgrupper, for å øke kunnskapen om elevers tekstproduksjon på norsk generelt og om psykolingvistiske betingelser for språkproduksjon mer generelt. Når det gjelder akkurat kontrasten mellom håndskriving og tastaturskriving, er jeg redd utviklingen er kommet så langt at det nå er vanskelig å foreta den type sammenlignende studie som jeg har gjort; elever skriver mindre og mindre for hånd, akkurat som andre medlemmer av samfunnet.

Jeg tror det er et potensial i utforsking av leksikalske variabler med tanke på automatisk eller halvautomatisk analyse av teksters kompleksitet, elevers modenhet og til beslektede formål. Jeg vil imøtese videre utforsking av TTR-baserte mål, men tror vi trenger supplerende variabler som mer valid representerer relevante tekstlige egenskaper enn det repetisjonsbaserte TTR-mål gjør.

Det er begrenset hva man kan finne ut om elevers skriving ved å studere produktene. Jeg tror vi kan få mer nyansert og kanskje mer nøyaktig innsikt i prosesser, ferdigheter, oppfatninger og holdninger ved å kombinere korpusbaserte studier med andre metoder, for eksempel med mer kvalitativ tekstanalyse, men også ved å studere skriveprosessens detaljer gjennom teknikker som tastelogging, blikksporing og høyttenkningsprotokoll, og gjennom intervjuer eller andre typer dialogbaserte dybdeundersøkelser av elevene.

Jeg har i løpet av prosessen noen ganger fått spørsmålet om hvorfor jeg ikke heller analyserer tekstene med tanke på oppbygning, innhold eller kvalitet. Jeg mener det er viktig at vi utvikler mer kunnskap om skriveprosesser og skriveverktøy også med tanke på tekststruktur, momentrikdom og tekstkvalitet, og at dette er viktige bestanddeler i en kunnskapsbasert skrivedidaktikk. Slike studier bør også kombineres med studier av språklige variabler slik at vi kan få vite mer om sammenhengen mellom språktrekk og kvalitet. I et slikt bilde utgjør denne avhandlingen en brikke i et større puslespill.

A. Analyser

A1. Trinnvis modellreduksjon

Leksikalske variabler:

```
lm.ordlengde1 <- lm(lexD$ordlengde~(kjønn+ferdighet+lengde+forskjell)^2)
lm.ordlengde2 <- step(lm.ordlengde1)
summary.aov(lm.ordlengde2)
lm.ordlengde3 <- update(lm.ordlengde2,~.-ferdighet:forskjell)
summary.aov(lm.ordlengde3)
lm.ordlengde4 <- step(lm.ordlengde3)
summary.aov(lm.ordlengde4)
gvlma(lm.ordlengde4)

lm.leksordF1 <- lm(lexD$leksordF ~ (kjønn+ferdighet+lengde+forskjell)^2)
lm.leksordF2 <- step(lm.leksordF1)
summary.aov(lm.leksordF2)
TukeyHSD(aov(lm.leksordF2))

lm.leksordkl1 <- lm(lexD$leksord.kl ~
(kjønn+ferdighet+lengde+forskjell)^2)
lm.leksordkl2 <- step(lm.leksordkl1)
summary.aov(lm.leksordkl2)
lm.leksordkl3 <- update(lm.leksordkl2,~.-kjønn:forskjell)
summary.aov(lm.leksordkl3)
lm.leksordkl4 <- step(lm.leksordkl3)
summary.aov(lm.leksordkl4)

lm.avislfi1 <- lm(lexD$avislfi~(kjønn+ferdighet+lengde+forskjell)^2)
lm.avislfi2 <- step(lm.avislfi1)
summary.aov(lm.avislfi2)
lm.avislfi3 <- update(lm.avislfi2,~.-forskjell)
summary.aov(lm.avislfi3)
summary.lm(lm.avislfi3)
gvlma(lm.avislfi3)

lm.elevlfi1 <- lm(lexD$elevlfi~(kjønn+ferdighet+lengde+forskjell)^2)
lm.elevlfi2 <- step(lm.elevlfi1)
summary.aov(lm.elevlfi2)
lm.elevlfi3 <- update(lm.elevlfi2,~.-kjønn:forskjell)
summary.aov(lm.elevlfi3)
lm.elevlfi4 <- update(lm.elevlfi3,~.-kjønn)
summary.aov(lm.elevlfi4)
summary.lm(lm.elevlfi4)
gvlma(lm.elevlfi4)

lm.leksordlengde1 <- lm(lexD$leksordlengde ~
(kjønn+ferdighet+lengde+forskjell)^2)
lm.leksordlengde2 <- step(lm.leksordlengde1)
summary.aov(lm.leksordlengde2)
lm.leksordlengde3 <- step(update(lm.leksordlengde2,~.-kjønn:lengde))
summary.aov(lm.leksordlengde3)
lm.leksordlengde4 <- step(update(lm.leksordlengde3,~.-kjønn:ferdighet))
summary.aov(lm.leksordlengde4)
lm.leksordlengde5 <- step(update(lm.leksordlengde4,~.-ferdighet:lengde))
summary.aov(lm.leksordlengde5)
summary.lm(lm.leksordlengde5)
```

```
gvlma(lm.leksordlengde5)

lm.log.TTR1 <- lm(lexD$log.TTR~(kjønn + ferdighet + lengde + forskjell)^2)
lm.log.TTR2 <- step(lm.log.TTR1)
summary.aov(lm.log.TTR2)
lm.log.TTR3 <- step(update(lm.log.TTR2,~.-ferdighet:lengde))
summary.aov(lm.log.TTR3)
summary.lm(lm.log.TTR3)
gvlma(lm.log.TTR3)

lm.OVIX1 <- lm(lexD$OVIX~(kjønn + ferdighet + lengde + forskjell)^2)
lm.OVIX2 <- step(lm.OVIX1)
summary.aov(lm.OVIX2)
lm.OVIX3 <- step(update(lm.OVIX2,~.-ferdighet:lengde))
summary.aov(lm.OVIX3)
summary.lm(lm.OVIX3)
gvlma(lm.OVIX3)

lexD$BrunetsW <- lex$BrunetsW[Tast] - lex$BrunetsW[Hånd]
lm.BrunetsW1 <- lm(lexD$BrunetsW~(kjønn+ferdighet+lengde+forskjell)^2)
lm.BrunetsW2 <- step(lm.BrunetsW1)
summary.aov(lm.BrunetsW2)
lm.BrunetsW3 <- step(update(lm.BrunetsW2,~.-ferdighet:lengde))
summary.aov(lm.BrunetsW3)
summary.lm(lm.BrunetsW3)
gvlma(lm.BrunetsW3)

lexD$BrunetsW.a0255 <- lex$BrunetsW.a0255[Tast] - lex$BrunetsW.a0255[Hånd]
lm.BrunetsW.a0255_1 <- lm(lexD$BrunetsW.a0255~(kjønn + ferdighet + lengde
+ forskjell)^2)
lm.BrunetsW.a0255_2 <- step(lm.BrunetsW.a0255_1)
summary.aov(lm.BrunetsW.a0255_2)
lm.BrunetsW.a0255_3 <- step(update(lm.BrunetsW.a0255_2,~.-
ferdighet:lengde))
summary.aov(lm.BrunetsW.a0255_3)
summary.lm(lm.BrunetsW.a0255_3)
gvlma(lm.BrunetsW.a0255_3)

lm.logTTR13_1 <- lm(lexD$log.TTR.13~(kjønn+ferdighet+lengde+forskjell)^2)
lm.logTTR13_2 <- step(lm.logTTR13_1)
summary.aov(lm.logTTR13_2)
lm.logTTR13_3 <- step(update(lm.logTTR13_2,~.-ferdighet:lengde))
summary.aov(lm.logTTR13_3)
summary.lm(lm.logTTR13_3)
gvlma(lm.logTTR13_3)

lexD$MOSTTR_LL.50 <- lex$MOSTTR_LL.50[Tast] - lex$MOSTTR_LL.50[Hånd]
lm.MOSTTR_LL50.1 <-
lm(lexD$MOSTTR_LL.50~(kjønn+ferdighet+lengde+forskjell)^2)
lm.MOSTTR_LL50.2 <- step(lm.MOSTTR_LL50.1)
summary.aov(lm.MOSTTR_LL50.2)
lm.MOSTTR_LL50.3 <- step(update(lm.MOSTTR_LL50.2, ~.-ferdighet:lengde))
summary.aov(lm.MOSTTR_LL50.3)
lm.MOSTTR_LL50.4 <- update(lm.MOSTTR_LL50.3, ~.-kjønn)
summary.aov(lm.MOSTTR_LL50.4)
summary.lm(lm.MOSTTR_LL50.4)
gvlma(lm.MOSTTR_LL50.4)

lexD$hapax.lemmaF <- lex$hapax.lemmaF[Tast] - lex$hapax.lemmaF[Hånd]
lm.hapax1 <- lm(lexD$hapax.lemmaF ~ (kjønn+ferdighet+lengde+forskjell)^2)
lm.hapax2 <- step(lm.hapax1)
```

```
summary.aov(lm.hapax2)
lm.hapax3 <- step(update(lm.hapax2, ~.-ferdighet:forskjell))
summary.aov(lm.hapax3)
lm.hapax4 <- update(lm.hapax3, ~.- ferdighet)
summary.aov(lm.hapax4)
summary.lm(lm.hapax4)
gvlma(lm.hapax4)
TukeyHSD(aov(lm.hapax4))

lexD$MAentropi.lem.100 <- lex$MAentropi.lem.100[Tast] -
lex$MAentropi.lem.100[Hånd]
lm.MAentropi100.1 <-
lm(lexD$MAentropi.lem.100~(kjønn+ferdighet+lengde+forskjell)^2)
lm.MAentropi100.2 <- step(lm.MAentropi100.1)
summary.aov(lm.MAentropi100.2)
lm.MAentropi100.3 <- step(update(lm.MAentropi100.2,~.-lengde:forskjell))
summary.aov(lm.MAentropi100.3)
summary.lm(lm.MAentropi100.3)
gvlma(lm.MAentropi100.3)
cohens.d(lexD$MAentropi.lem.100[gutt],lexD$MAentropi.lem.100[jente])

lexD$MOSentropi.lem.100 <- lex$MOSentropi.lem.100[Tast] -
lex$MOSentropi.lem.100[Hånd]
lm.MOSentropi100.1 <-
lm(lexD$MOSentropi.lem.100~(kjønn+ferdighet+lengde+forskjell)^2)
lm.MOSentropi100.2 <- step(lm.MOSentropi100.1)
summary.aov(lm.MOSentropi100.2)
summary.lm(lm.MOSentropi100.2)
gvlma(lm.MOSentropi100.2)
TukeyHSD(aov(lm.MOSentropi100.2))
lm.MOSentropi100.3 <- step(update(lm.MOSentropi100.2,~.-lengde:forskjell))
summary.aov(lm.MOSentropi100.3)
summary.lm(lm.MOSentropi100.3)
```

Syntaktiske variabler:

```
lm.TEL1 <- lm(synD$lTEL~(kjønn+ferdighet+lengde+forskjell)^2)
lm.TEL2 <- step(lm.TEL1)
summary.aov(lm.TEL2)
lm.TEL3 <- step(update(lm.TEL2, ~.- ferdighet:lengde))
summary.aov(lm.TEL3)
summary.lm(lm.TEL3)
cohens.d(syn$lTEL[Tast], syn$lTEL[Hånd])

lm.TELleks1 <- lm(synD$lTEL.leks~(kjønn+ferdighet+lengde+forskjell)^2)
lm.TELleks2 <- step(lm.TELleks1)
summary.aov(lm.TELleks2)
lm.TELleks3 <- step(update(lm.TELleks2,~.-lengde:forskjell))
summary.aov(lm.TELleks3)
summary.lm(lm.TELleks3)
#cohens.d(syn$lTEL.leks[Tast], syn$lTEL.leks[Hånd])

lm.klL1 <- lm (synD$klL~(kjønn+ferdighet+lengde+forskjell)^2)
lm.klL2 <- step(lm.klL1)
summary.aov(lm.klL2)
lm.klL3 <- step(update(lm.klL2,~.-kjønn:forskjell))
summary.aov(lm.klL3)
lm.klL4 <- step(update(lm.klL3,~.-forskjell))
summary.aov(lm.klL4)
lm.klL5 <- step(update(lm.klL4,~.-lengde))
summary.aov(lm.klL5)
```

```
summary.lm(lm.klL5)
t.test(syn$klL~M, paired=T)

lm.klLleks1 <- lm (synD$klL.leks~(kjønn+ferdighet+lengde+forskjell)^2)
lm.klLleks2 <- step(lm.klLleks1)
summary.aov(lm.klLleks2)
lm.klLleks3 <- step(update(lm.klLleks2,~.-kjønn:forskjell))
summary.aov(lm.klLleks3)
summary.lm(lm.klLleks3)
gvlma(lm.klLleks3)
TukeyHSD(aov(lm.klLleks3))
lm.klLleks4 <- step(update(lm.klLleks3,~.-ferdighet:forskjell))
summary.aov(lm.klLleks4)
lm.klLleks5 <- step(update(lm.klLleks4,~.-forskjell))
summary.aov(lm.klLleks5)
summary.lm(lm.klLleks5)

lm.subklL3F.1 <- lm(synD$subklL_3F~(kjønn+ferdighet+lengde+forskjell)^2)
lm.subklL3F.2 <- step(lm.subklL3F.1)
summary.aov(lm.subklL3F.2)
summary.lm(lm.subklL3F.2)
gvlma(lm.subklL3F.2)
TukeyHSD(aov(lm.subklL3F.2))

lm.prepkl1 <- lm(posD$prep.kl~(kjønn+ferdighet+lengde+forskjell)^2)
lm.prepkl2 <- step(lm.prepkl1)
summary.aov(lm.prepkl2)
lm.prepkl3 <- step(update(lm.prepkl2,~.-kjønn))
summary.aov(lm.prepkl3)
lm.prepkl4 <- step(update(lm.prepkl3,~.-lengde:forskjell))
summary.aov(lm.prepkl4)
lm.prepkl5 <- step(update(lm.prepkl4,~.-forskjell))
summary.aov(lm.prepkl5)
summary.lm(lm.prepkl5)
gvlma(lm.prepkl5)

lm.klausadvkl1 <-
lm(synD$klaus.adv.kl~(kjønn+ferdighet+lengde+forskjell)^2)
lm.klausadvkl2 <- step(lm.klausadvkl1)
summary.aov(lm.klausadvkl2)
lm.klausadvkl3 <- step(update(lm.klausadvkl2,~.-kjønn:forskjell))
summary.aov(lm.klausadvkl3)
summary.lm(lm.klausadvkl3)
gvlma(lm.klausadvkl3)

lm.subklte1 <- lm(synD$subkl.te~(kjønn+ferdighet+lengde+forskjell)^2)
lm.subklte2 <- step(lm.subklte1)
summary.aov(lm.subklte2)
lm.subklte3 <- update(lm.subklte2,~.-forskjell)
summary.aov(lm.subklte3)
summary.lm(lm.subklte3)

lm.subklF1 <- lm(synD$subklF~(kjønn+ferdighet+lengde+forskjell)^2)
lm.subklF2 <- step(lm.subklF1)
summary.aov(lm.subklF2)
lm.subklF3 <- update(lm.subklF2,~.-forskjell)
summary.aov(lm.subklF3)
summary.lm(lm.subklF3)

lm.TEind.1ffF1 <-
lm(synD$logit.TEind.1ffF~(kjønn+ferdighet+lengde+forskjell)^2)
```

```
lm.TEind.1ffF2 <- step(lm.TEind.1ffF1)
summary.aov(lm.TEind.1ffF2)
lm.TEind.1ffF3 <- step(update(lm.TEind.1ffF2,~.-ferdighet:forskjell))
summary.aov(lm.TEind.1ffF3)
summary.lm(lm.TEind.1ffF3)
gvlma(lm.TEind.1ffF3)
TukeyHSD(aov(lm.TEind.1ffF3))

lm.AAred1 <- lm(posD$AA.redF ~ (kjønn + ferdighet + lengde + forskjell)^2)
lm.AAred2 <- step(lm.AAred1)
summary.aov(lm.AAred2)
lm.AAred3 <- update(lm.AAred2,~.-kjønn:forskjell)
summary.aov(lm.AAred3)
lm.AAred4 <- update(lm.AAred3,~.-forskjell)
summary.aov(lm.AAred4)
lm.AAred5 <- update(lm.AAred4,~.-kjønn)
summary.aov(lm.AAred5)
summary.lm(lm.AAred5)
gvlma(lm.AAred1)
gvlma(lm.AAred2)
gvlma(lm.AAred3)
gvlma(lm.AAred4)
gvlma(lm.AAred5)
t.test(pos$AA.redF[Tast], pos$AA.redF[Hånd], paired=T)
shapiro.test(posD$AA.redF)

pos$AA.red.subst <- pos$AA.red / pos$subst
posD$AA.red.subst <- pos$AA.red.subst[Tast] - pos$AA.red.subst[Hånd]
lm.AA.red.subst1 <- lm(posD$AA.red.subst ~ (kjønn + ferdighet + lengde +
forskjell)^2)
lm.AA.red.subst2 <- step(lm.AA.red.subst1)
summary.aov(lm.AA.red.subst2)
lm.AA.red.subst3 <- update(lm.AA.red.subst2,~.-kjønn:ferdighet)
summary.aov(lm.AA.red.subst3)
lm.AA.red.subst4 <- update(lm.AA.red.subst3,~.-kjønn:forskjell)
summary.aov(lm.AA.red.subst4)
lm.AA.red.subst5 <- step(lm.AA.red.subst4)
summary.aov(lm.AA.red.subst5)
summary.lm(lm.AA.red.subst5)
gvlma(lm.AA.red.subst5)

pos$AA.red.kl    <- pos$AA.red / syn$kl
posD$AA.red.kl <- pos$AA.red.kl[Tast] - pos$AA.red.kl[Hånd]
lm.AA.red.kl1 <- lm(posD$AA.red.kl ~ (kjønn + ferdighet + lengde +
forskjell)^2)
lm.AA.red.kl2 <- step(lm.AA.red.kl1)
summary.aov(lm.AA.red.kl2)
AIC(lm.AA.red.kl1); AIC(lm.AA.red.kl2)
lm.AA.red.kl3 <- update(lm.AA.red.kl2,~.-kjønn:forskjell)
summary.aov(lm.AA.red.kl3)
AIC(lm.AA.red.kl3)
lm.AA.red.kl4 <- step(lm.AA.red.kl3)
summary.aov(lm.AA.red.kl4)
summary.lm(lm.AA.red.kl4)
```

Prinsipalkomponenter:

```
lm.pc1 <- step(lm(pcaD.df$pc1~(kjønn+ferdighet+lengde+forskjell)^2))
lm.pc2 <- step(lm(pcaD.df$pc2~(kjønn+ferdighet+lengde+forskjell)^2))
lm.pc3 <- step(lm(pcaD.df$pc3~(kjønn+ferdighet+lengde+forskjell)^2))
```

```
lm.pc4 <- step(lm(pcaD.df$pc4~(kjønn+ferdighet+lengde+forskjell)^2))

lm.pc12 <- step(lm(pc12~(kjønn+ferdighet+lengde+forskjell)^2))

summary.aov(lm.pc1)
summary.aov(lm.pc2)
summary.aov(lm.pc3)
summary.aov(lm.pc4)

lm.pc1.2 <- update(lm.pc1,~.-kjønn:lengde)
summary.aov(lm.pc1.2)
lm.pc1.3 <- update(lm.pc1.2,~.-kjønn:ferdighet)
summary.aov(lm.pc1.3)
lm.pc1.4 <- update(lm.pc1.3,~.-ferdighet:forskjell)
summary.aov(lm.pc1.4)
lm.pc1.5 <- update(lm.pc1.4,~.-ferdighet:lengde)
summary.aov(lm.pc1.5)
lm.pc1.6 <- update(lm.pc1.5,~.-ferdighet)
summary.aov(lm.pc1.6)
lm.pc1.7 <- update(lm.pc1.6,~.-kjønn)
summary.aov(lm.pc1.7)
lm.pc1.8 <- update(lm.pc1.7,~.-lengde)
summary.aov(lm.pc1.8)
gvlma(lm.pc1.8)    # OK

lm.pc2.2 <- update(lm.pc2,~.-ferdighet:forskjell)
summary.aov(lm.pc2.2)
lm.pc2.3 <- step(lm.pc2.2)
summary.aov(lm.pc2.3)
gvlma(lm.pc2.3)    # OK

lm.pc3.2 <- update(lm.pc3,~.-ferdighet:forskjell)
summary.aov(lm.pc3.2)
lm.pc3.3 <- update(lm.pc3.2,~.-ferdighet:lengde)
summary.aov(lm.pc3.3)
lm.pc3.4 <- update(lm.pc3.3,~.-ferdighet)
summary.aov(lm.pc3.4)
summary.lm(lm.pc3.4)
gvlma(lm.pc3.4)    # OK

lm.pc4.2 <- update(lm.pc4,~.-lengde:forskjell)
summary.aov(lm.pc4.2)
lm.pc4.3 <- step(lm.pc4.2)
summary.aov(lm.pc4.3)
lm.pc4.4 <- update(lm.pc4.3,~.-lengde)
summary(lm.pc4.4)

summary.aov(lm.pc1.8)
summary.aov(lm.pc2.3)
summary.aov(lm.pc3.4)
summary(lm.pc4.4)

TukeyHSD(aov(lm.pc3.4))
```

A2. Anova-tabeller

Anova-tabeller for de 5 + 8 viktigste variablene.

```
lm(formula = lexD$ordlengde ~ kjønn)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
kjønn        1 0.2037  0.2037   6.789 0.0116 *
Residuals   58 1.7400  0.0300
---
Multiple R-squared:  0.1048,    Adjusted R-squared:  0.08936
F-statistic: 6.789 on 1 and 58 DF,  p-value: 0.01163

lm(formula = lexD$leksordF ~ ferdighet + forskjell + ferdighet:forskjell)
                   Df  Sum Sq  Mean Sq F value Pr(>F)
ferdighet           1 0.00046 0.000464   0.292 0.5913
forskjell           1 0.00527 0.005269   3.312 0.0741 .
ferdighet:forskjell  1 0.00883 0.008830   5.550 0.0220 *
Residuals           56 0.08910 0.001591
---
Multiple R-squared:  0.1405,    Adjusted R-squared:  0.09444
F-statistic: 3.051 on 3 and 56 DF,  p-value: 0.03588

lm(formula = lexD$leksordlengde ~ kjønn)
Df Sum Sq Mean Sq F value  Pr(>F)
kjønn        1 0,5338 0,53375  4,4251 0,03976 *
Residuals   58 6,9959 0,12062
----
Multiple R-squared:  0.07089,   Adjusted R-squared:  0.05487
F-statistic: 4.425 on 1 and 58 DF,  p-value: 0.03976

lm(formula = lexD$log.TTR.13 ~ kjønn + forskjell)
            Df  Sum Sq  Mean Sq F value   Pr(>F)
kjønn        1 0.01584 0.015841   8.512 0.005041 **
forskjell    1 0.02648 0.026481  14.229 0.000387 ***
Residuals   57 0.10608 0.001861
---
Multiple R-squared:  0.2852,    Adjusted R-squared:  0.2601
F-statistic: 11.37 on 2 and 57 DF,  p-value: 6.993e-05

lm(formula = lexD$MOSTTR_LL.50 ~ forskjell)
            Df Sum Sq Mean Sq F value Pr(>F)
forskjell    1 0.0666 0.06659   9.267 0.0035 **
Residuals   58 0.4168 0.00719
----
Multiple R-squared:  0.1378,    Adjusted R-squared:  0.1229
F-statistic: 9.267 on 1 and 58 DF,  p-value: 0.003504

lm(formula = synD$lTEL ~ 1)
(Intercept)  0.06533    0.02369   2.758  0.00773 **
---
Residual standard error: 0.1835 on 59 degrees of freedom

lm(formula = synD$klL ~ 1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04796    0.10673   0.449    0.655
---
Residual standard error: 0.8267 on 59 degrees of freedom

lm(synD$subklL_3F~(kjønn+ferdighet+lengde+forskjell)^2)
                Df    Sum Sq   Mean Sq F value Pr(>F)
kjønn            1 0.0000065 6.460e-06   0.210 0.6488
lengde           1 0.0002028 2.028e-04   6.589 0.0130 *
forskjell        1 0.0000172 1.718e-05   0.558 0.4583
kjønn:forskjell  1 0.0001409 1.409e-04   4.577 0.0369 *
Residuals       55 0.0016932 3.078e-05
---
```

```
Multiple R-squared:  0.1783,    Adjusted R-squared:  0.1185
F-statistic: 2.983 on 4 and 55 DF,  p-value: 0.02666

lm(formula = posD$prep.kl ~ lengde)
            Df Sum Sq Mean Sq F value Pr(>F)
lengde       1 0.2246 0.22460   6.045  0.017 *
Residuals   58 2.1550 0.03716
---
Multiple R-squared:  0.09439,   Adjusted R-squared:  0.07877
F-statistic: 6.045 on 1 and 58 DF,  p-value: 0.01695

lm(formula = synD$klaus.adv.kl ~ forskjell)
            Df  Sum Sq Mean Sq F value  Pr(>F)
forskjell    1 0.04878 0.04878   10.09 0.00239 **
Residuals   58 0.28053 0.00484
---
Multiple R-squared:  0.1481,    Adjusted R-squared:  0.1334
F-statistic: 10.09 on 1 and 58 DF,  p-value: 0.002394

lm(formula = synD$subkl.te ~ 1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.11036    0.04866   2.268    0.027 *
---
Residual standard error: 0.3769 on 59 degrees of freedom

lm(formula = synD$logit.TEind.1ffF ~ kjønn + forskjell + kjønn:forskjell)
                Df Sum Sq Mean Sq F value Pr(>F)
kjønn            1  1.568  1.5679   5.351 0.0244 *
forskjell        1  0.940  0.9398   3.207 0.0787 .
kjønn:forskjell  1  1.410  1.4098   4.811 0.0324 *
Residuals       56 16.409  0.2930
---
Multiple R-squared:  0.1927,    Adjusted R-squared:  0.1495
F-statistic: 4.456 on 3 and 56 DF,  p-value: 0.007069

lm(formula = posD$AA.redF ~ 1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001436   0.001275   1.126    0.265
---
Residual standard error: 0.009877 on 59 degrees of freedom
```

A3. Anova-tabeller (andre)

Her gjengis resultater fra anova-analyser som ikke er gjengitt i løpeteksten.

Anova-analyse for OVIX (s. 184). `Gvlma` (7.2.2.4) viser at premissene for anova er oppfylt (se appendiks A4).

```
            Df Sum Sq Mean Sq F value Pr(>F)
kjønn        1  238.5  238.51   8.752 0.0045 **
forskjell    1  138.1  138.10   5.067 0.0283 *
Residuals   57 1553.3   27.25
---
Multiple R-squared:  0.1951,    Adjusted R-squared:  0.1669
F-statistic:  6.91 on 2 and 57 DF,  p-value: 0.002056
```

Anova-analyse for Brunets W (s. 187). `Gvlma` (7.2.2.4) viser at premissene for anova er oppfylt (se appendiks A4). Ettersom tekstverdiene av Brunets W i motsetning til

differanseverdiene er høyreskjeve, bør resultatene av anova-analyse på differanseverdiene behandles med en viss varsomhet.

```
            Df Sum Sq Mean Sq F value  Pr(>F)
kjønn        1  1.091  1.0914  11.334 0.00137 **
forskjell    1  0.402  0.4021   4.176 0.04563 *
Residuals   57  5.489  0.0963
---
Multiple R-squared:  0.2139,    Adjusted R-squared:  0.1863
F-statistic: 7.755 on 2 and 57 DF,  p-value: 0.00105
```

Anova-analyse for Brunets W med a = 0,255 (s. 191).

```
lm(formula = lexD$BrunetsW.a0255 ~ kjønn + forskjell)
            Df Sum Sq Mean Sq F value   Pr(>F)
kjønn        1 0.4800  0.4800   9.496 0.003169 **
forskjell    1 0.7984  0.7984  15.795 0.000201 ***
Residuals   57 2.8813  0.0505
---
Multiple R-squared:  0.3073,    Adjusted R-squared:  0.283
F-statistic: 12.65 on 2 and 57 DF,  p-value: 2.85e-05
```

Anova-analyse for MOS-entropi$_{W=100}$ (s. 237), som resulterer i den ikke-signifikante minimale modellen under ($F \approx 2{,}76$, $p \approx 0{,}0504$). `Gvlma` (7.2.2.4) viser at premissene for anova er oppfylt (se appendiks A4). Tukeys HSD-test gir heller ingen signifikante interaksjoner (se appendiks A5). (Ettersom modellen ikke er signifikant, er det egentlig uansett ikke behov for å bruke noen post-hoc test.)

```
> summary.aov(lm.MOSentropi100.2)
                Df    Sum Sq   Mean Sq F value  Pr(>F)
lengde           1 0.0000056 5.580e-06   0.183 0.67072
forskjell        1 0.0000031 3.110e-06   0.102 0.75077
lengde:forskjell 1 0.0002444 2.444e-04   8.004 0.00647 **
Residuals       56 0.0017096 3.053e-05
---
Multiple R-squared:  0.1289,    Adjusted R-squared:  0.08226
F-statistic: 2.763 on 3 and 56 DF,  p-value: 0.05039
```

Anova-analyse for antall leksikalske ord per t-enhet (s. 253), som resulterer i ikke-signifikant nullmodell:

```
> summary.aov(lm.TELleks3)
            Df Sum Sq Mean Sq F value Pr(>F)
Residuals   59  1.874 0.03176
> summary.lm(lm.TELleks3)

Call:
lm(formula = synD$lTEL.leks ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40131 -0.06909  0.00995  0.10525  0.44387

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04369    0.02301   1.899   0.0625 .
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1782 on 59 degrees of freedom
```

Anova-analyse for gjennomsnittlig antall leksikalske ord per klausus (s. 258), som resulterer i den ikke-signifikante ($F \approx 2{,}60$, $p \approx 0{,}061$) minimale modellen nedenfor. `Gvlma` (7.2.2.4) viser at premissene for anova er oppfylt (se appendiks A4). Tukeys HSD-test rapporterer en signifikant forskjell mellom sterke elever med liten tekstlengdeforskjell og sterke elever med stor tekstlengdeforskjell ($p \approx 0{,}041$), men ettersom modellen som helhet ikke er signifikant, skal dette ikke tas hensyn til. (Se appendiks A5).

```
                    Df Sum Sq Mean Sq F value Pr(>F)
ferdighet            1  0.065  0.0646   0.287 0.5945
forskjell            1  0.702  0.7021   3.115 0.0830 .
ferdighet:forskjell  1  0.989  0.9894   4.390 0.0407 *
Residuals           56 12.622  0.2254
---
Multiple R-squared:  0.1221,    Adjusted R-squared:  0.07511
F-statistic: 2.597 on 3 and 56 DF,  p-value: 0.0613
```

Videre reduksjon av modellen gir som resultat en ikke-signifikant nullmodell:

```
            Df Sum Sq Mean Sq F value Pr(>F)
Residuals   59  14.38  0.2437

Call:
lm(formula = synD$klL.leks ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.16311 -0.33171  0.05067  0.35584  0.98776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03590    0.06373  -0.563    0.575

Residual standard error: 0.4937 on 59 degrees of freedom
```

Anova-analyse for antall attributive adjektiver per substantiv, (s. 298):

```
            Df  Sum Sq  Mean Sq F value Pr(>F)
forskjell    1 0.01276 0.012762   5.866 0.0186 *
Residuals   58 0.12618 0.002176
---
Multiple R-squared:  0.09185,   Adjusted R-squared:  0.07619
F-statistic: 5.866 on 1 and 58 DF,  p-value: 0.01858
```

`Gvlma` (7.2.2.4) rapporterer at premissene for anova er oppfylt (se appendiks A4).

A4. Resultater fra gvlma

Anova-analyse av ordlengde (s. 138).

```
Call:
lm(formula = lexD$ordlengde ~ kjønn)
```

```
Coefficients:
(Intercept)       kjønnJ
    0.04489     -0.11653


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm(lexD$ordlengde ~ kjønn))

                      Value p-value                Decision
Global Stat         3.005e+00  0.5570 Assumptions acceptable.
Skewness            2.460e+00  0.1168 Assumptions acceptable.
Kurtosis            5.407e-01  0.4621 Assumptions acceptable.
Link Function      -7.125e-15  1.0000 Assumptions acceptable.
Heteroscedasticity  4.331e-03  0.9475 Assumptions acceptable.
```

Anova-analyse av leksikalsk tetthet (s. 150).

```
Call:
lm(formula = lexD$leksordF ~ ferdighet + forskjell + ferdighet:forskjell)

Coefficients:
             (Intercept)                ferdighetS
forskjellstor  ferdighetS:forskjellstor
               -0.013885                  0.029823
0.005519                -0.048524


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.leksordF2)

                      Value p-value                Decision
Global Stat         2.988e+00  0.5599 Assumptions acceptable.
Skewness            1.616e-01  0.6877 Assumptions acceptable.
Kurtosis            7.199e-01  0.3962 Assumptions acceptable.
Link Function      -1.927e-14  1.0000 Assumptions acceptable.
Heteroscedasticity  2.106e+00  0.1467 Assumptions acceptable.
```

Anova-analyse av leksikalske ord per klausus (s. 338).

```
> gvlma(lm.leksordkl4)

Call:
lm(formula = lexD$leksord.kl ~ ferdighet + forskjell +
ferdighet:forskjell)

Coefficients:
             (Intercept)                ferdighetS
forskjellstor
                -0.08896                   0.32246
0.04048
ferdighetS:forskjellstor
                -0.51365
```

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.leksordkl4)

                     Value p-value                Decision
Global Stat        4.592e+00  0.3318 Assumptions acceptable.
Skewness           1.833e+00  0.1758 Assumptions acceptable.
Kurtosis           1.934e+00  0.1643 Assumptions acceptable.
Link Function      1.742e-18  1.0000 Assumptions acceptable.
Heteroscedasticity 8.250e-01  0.3637 Assumptions acceptable.
```

Anova-analyse av leksikalsk sofistikerthet (avisbasert *lfi*) (s. 157).

```
Call:
lm(formula = lexD$avislfi ~ lengde)

Coefficients:
(Intercept)   lengdelang
   0.002126    -0.007625


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.avislfi3)

                       Value p-value                   Decision
Global Stat         5.212e+00 0.26622    Assumptions acceptable.
Skewness            3.975e+00 0.04617 Assumptions NOT satisfied!
Kurtosis            7.734e-02 0.78094    Assumptions acceptable.
Link Function      -1.297e-16 1.00000    Assumptions acceptable.
Heteroscedasticity  1.159e+00 0.28159    Assumptions acceptable.
```

Anova-analyse av leksikalsk originalitet (*lfi* basert på korpuset av elevtekster) (s.158).

```
Call:
lm(formula = lexD$elevlfi ~ forskjell)

Coefficients:
  (Intercept)  forskjellstor
     -0.01173        0.01948


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.elevlfi4)

                       Value p-value                Decision
Global Stat         2.019e+00  0.7323 Assumptions acceptable.
Skewness            9.161e-01  0.3385 Assumptions acceptable.
Kurtosis            1.729e-01  0.6776 Assumptions acceptable.
Link Function      -2.961e-16  1.0000 Assumptions acceptable.
```

```
Heteroscedasticity  9.299e-01  0.3349 Assumptions acceptable.
```

Anova-analyse av ordlengde i leksikalske ord (s. 163).

```
Call:
lm(formula = lexD$leksordlengde ~ kjønn)

Coefficients:
(Intercept)       kjønnJ
    0.06152     -0.18864


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.leksordlengde5)

                     Value p-value                Decision
Global Stat        2.087e+00  0.7197 Assumptions acceptable.
Skewness           1.864e-02  0.8914 Assumptions acceptable.
Kurtosis           6.195e-01  0.4312 Assumptions acceptable.
Link Function      3.556e-15  1.0000 Assumptions acceptable.
Heteroscedasticity 1.449e+00  0.2287 Assumptions acceptable.
```

Anova-analyse av log-TTR (s. 180).

```
Call:
lm(formula = lexD$log.TTR ~ kjønn + forskjell)

Coefficients:
  (Intercept)         kjønnJ  forskjellstor
    -0.008383      -0.033578       0.022221


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.log.TTR3)

                    Value p-value                Decision
Global Stat        2.6364  0.6204 Assumptions acceptable.
Skewness           0.1113  0.7387 Assumptions acceptable.
Kurtosis           0.6656  0.4146 Assumptions acceptable.
Link Function      0.3452  0.5568 Assumptions acceptable.
Heteroscedasticity 1.5143  0.2185 Assumptions acceptable.
```

Anova-analyse av OVIX (s. 336).

```
Call:
lm(formula = lexD$OVIX ~ kjønn + forskjell)

Coefficients:
  (Intercept)         kjønnJ  forskjellstor
       -0.942         -3.988          3.034


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
```

```
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.OVIX3)

                     Value p-value                Decision
Global Stat        2.668784  0.6147 Assumptions acceptable.
Skewness           0.001594  0.9681 Assumptions acceptable.
Kurtosis           0.799389  0.3713 Assumptions acceptable.
Link Function      0.146811  0.7016 Assumptions acceptable.
Heteroscedasticity 1.720989  0.1896 Assumptions acceptable.
```

Anova-analyse av Brunets $W_{a=0,22}$ (s. 336).

```
Call:
lm(formula = lexD$BrunetsW ~ kjønn + forskjell)

Coefficients:
  (Intercept)         kjønnJ  forskjellstor
      0.06172        0.26974       -0.16373


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.BrunetsW3)

                      Value p-value                Decision
Global Stat        2.6724  0.6140 Assumptions acceptable.
Skewness           0.3794  0.5379 Assumptions acceptable.
Kurtosis           0.8041  0.3699 Assumptions acceptable.
Link Function      0.5371  0.4636 Assumptions acceptable.
Heteroscedasticity 0.9518  0.3293 Assumptions acceptable.
```

Anova-analyse av Brunets $W_{a=0.255}$ (s. 191).

```
Call:
lm(formula = lexD$BrunetsW.a0255 ~ kjønn + forskjell)

Coefficients:
  (Intercept)         kjønnJ  forskjellstor
      0.04376        0.17889       -0.23071


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.BrunetsW.a0255_3)

                      Value p-value                Decision
Global Stat        2.9003  0.5746 Assumptions acceptable.
Skewness           0.6253  0.4291 Assumptions acceptable.
Kurtosis           0.9674  0.3253 Assumptions acceptable.
Link Function      0.6474  0.4210 Assumptions acceptable.
Heteroscedasticity 0.6602  0.4165 Assumptions acceptable.
```

Anova-analyse av justert log-$TTR_{1,3}$ (s. 200).

```
Call:
lm(formula = lexD$log.TTR.13 ~ kjønn + forskjell)

Coefficients:
  (Intercept)         kjønnJ  forskjellstor
    -0.008157      -0.032497       0.042016


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.logTTR13_3)

                     Value p-value                Decision
Global Stat        2.192173  0.7005 Assumptions acceptable.
Skewness           0.009644  0.9218 Assumptions acceptable.
Kurtosis           0.565171  0.4522 Assumptions acceptable.
Link Function      0.368102  0.5440 Assumptions acceptable.
Heteroscedasticity 1.249256  0.2637 Assumptions acceptable.
```

Anova-analyse av MOSTTR-$LL_{W=50}$ (s. 223).

```
Call:
lm(formula = lexD$MOSTTR_LL.50 ~ forskjell)

Coefficients:
  (Intercept)  forskjellstor
     -0.04699        0.06663


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.MOSTTR_LL50.4)

                       Value p-value                Decision
Global Stat         1.602e+00  0.8084 Assumptions acceptable.
Skewness            4.309e-01  0.5115 Assumptions acceptable.
Kurtosis            8.002e-01  0.3710 Assumptions acceptable.
Link Function      -2.743e-16  1.0000 Assumptions acceptable.
Heteroscedasticity  3.712e-01  0.5423 Assumptions acceptable.
```

Anova-analyse av hapax legomena per ordtyper (s. 228).

```
Call:
lm(formula = lexD$hapax.lemmaF ~ kjønn + forskjell + kjønn:forskjell)

Coefficients:
         (Intercept)                kjønnJ         forskjellstor
kjønnJ:forskjellstor
             0.01972             -0.06366              -0.05387
0.06430
```

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.hapax4)

                     Value p-value                Decision
Global Stat        8.563e-01  0.9307 Assumptions acceptable.
Skewness           6.660e-01  0.4145 Assumptions acceptable.
Kurtosis           1.282e-01  0.7203 Assumptions acceptable.
Link Function      3.284e-13  1.0000 Assumptions acceptable.
Heteroscedasticity 6.206e-02  0.8033 Assumptions acceptable.
```

Anova-analyse av MA-entropi$_{W=100}$ (s. 236).

```
Call:
lm(formula = lexD$MAentropi.lem.100 ~ kjønn)

Coefficients:
(Intercept)       kjønnJ
   0.001262    -0.003466


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.MAentropi100.3)

                     Value p-value                Decision
Global Stat        7.641e-01  0.9432 Assumptions acceptable.
Skewness           2.473e-01  0.6190 Assumptions acceptable.
Kurtosis           5.168e-01  0.4722 Assumptions acceptable.
Link Function      2.467e-14  1.0000 Assumptions acceptable.
Heteroscedasticity 2.433e-05  0.9961 Assumptions acceptable.
```

Anova-analyse av MOS-entropi$_{W=100}$ (s. 337).

```
Call:
lm(formula = lexD$MOSentropi.lem.100 ~ lengde + forskjell +
lengde:forskjell)

Coefficients:
             (Intercept)                lengdelang
forskjellstor
               -0.001764                  0.004636
0.004584
lengdelang:forskjellstor
               -0.008239


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.MOSentropi100.2)

                     Value p-value                Decision
```

```
Global Stat        1.081e+00  0.8972 Assumptions acceptable.
Skewness           1.073e+00  0.3002 Assumptions acceptable.
Kurtosis           2.971e-03  0.9565 Assumptions acceptable.
Link Function      2.999e-15  1.0000 Assumptions acceptable.
Heteroscedasticity 5.018e-03  0.9435 Assumptions acceptable.
```

Anova-analyse av gjennomsnittlig antall leksikalske ord per t-enhet resulterte i nullmodell (s. 253). Shapiro-Wilks normalitetstest rapporterer normal distribusjon i differanseverdiene av de logaritmetransformerte verdiene, $W \approx 0{,}065$. $p \approx 0{,}77$.

Anova-analyse av gjennomsnittlig antall leksikalske ord per klausus (s. 338).

```
Call:
lm(formula = synD$klL.leks ~ ferdighet + forskjell + ferdighet:forskjell)

Coefficients:
             (Intercept)                 ferdighetS
forskjellstor  ferdighetS:forskjellstor
                -0.08896                    0.32246
0.04048                  -0.51365


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.klLleks3)

                      Value p-value                Decision
Global Stat        4.592e+00  0.3318 Assumptions acceptable.
Skewness           1.833e+00  0.1758 Assumptions acceptable.
Kurtosis           1.934e+00  0.1643 Assumptions acceptable.
Link Function      1.742e-18  1.0000 Assumptions acceptable.
Heteroscedasticity 8.250e-01  0.3637 Assumptions acceptable.
```

Anova-analyse av frekvens av korte subklaususer (s. 263).

```
Call:
lm(formula = synD$subklL_3F ~ kjønn + lengde + forskjell +
kjønn:forskjell)

Coefficients:
         (Intercept)                kjønnJ            lengdelang
forskjellstor  kjønnJ:forskjellstor
            0.004244             -0.003483             -0.003680
-0.004121              0.006144


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.subklL3F.2)

                     Value p-value                   Decision
Global Stat        5.2192 0.26554    Assumptions acceptable.
Skewness           4.2455 0.03935 Assumptions NOT satisfied!
Kurtosis           0.1463 0.70207    Assumptions acceptable.
```

```
Link Function      0.2410 0.62346    Assumptions acceptable.
Heteroscedasticity 0.5863 0.44385    Assumptions acceptable.
```

Anova-analyse av andel korte subklaususer (s. 264).

```
Call:
lm(formula = synD$subklL_3.subkl ~ ferdighet + forskjell +
ferdighet:forskjell)

Coefficients:
             (Intercept)                 ferdighetS
forskjellstor  ferdighetS:forskjellstor
                 0.17097                     0.20347
0.04932                  -1.08914


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.subklL3subkl.3)

                      Value p-value                Decision
Global Stat        4.412e-01  0.9790 Assumptions acceptable.
Skewness           1.597e-01  0.6894 Assumptions acceptable.
Kurtosis           1.123e-01  0.7375 Assumptions acceptable.
Link Function      1.791e-15  1.0000 Assumptions acceptable.
Heteroscedasticity 1.691e-01  0.6809 Assumptions acceptable.
```

Anova-analyse av antall korte subklaususer per t-enhet (s. 264).

```
Call:
lm(formula = synD$subklL_3.te ~ ferdighet)

Coefficients:
(Intercept)   ferdighetS
    0.03170     -0.04333


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.subklL3te.5)

                       Value p-value                Decision
Global Stat         2.955e+00  0.5654 Assumptions acceptable.
Skewness            2.275e+00  0.1315 Assumptions acceptable.
Kurtosis            5.738e-01  0.4487 Assumptions acceptable.
Link Function      -5.329e-16  1.0000 Assumptions acceptable.
Heteroscedasticity  1.062e-01  0.7445 Assumptions acceptable.
```

Anova-analyse av antall preposisjoner per klausus (s. 274).

```
Call:
lm(formula = posD$prep.kl ~ lengde)

Coefficients:
(Intercept)   lengdelang
```

```
   -0.08758      0.12237


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.prepkl5)

                       Value p-value                Decision
Global Stat         2.551e+00  0.6356 Assumptions acceptable.
Skewness            8.461e-02  0.7711 Assumptions acceptable.
Kurtosis            2.299e-02  0.8795 Assumptions acceptable.
Link Function      -3.553e-16  1.0000 Assumptions acceptable.
Heteroscedasticity  2.443e+00  0.1180 Assumptions acceptable.
```

Anova-analyse av antall subklaususer per klausus (s. 279).

```
Call:
lm(formula = synD$klaus.adv.kl ~ forskjell)

Coefficients:
  (Intercept)  forskjellstor
     -0.02083        0.05703


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.klausadvkl3)

                       Value p-value                Decision
Global Stat         3.316e+00  0.5063 Assumptions acceptable.
Skewness            1.440e+00  0.2301 Assumptions acceptable.
Kurtosis            3.758e-01  0.5398 Assumptions acceptable.
Link Function      -1.234e-17  1.0000 Assumptions acceptable.
Heteroscedasticity  1.500e+00  0.2206 Assumptions acceptable.
```

Anova-analyse av andel t-enheter med kort forfelt (s. 292).

```
Call:
lm(formula = synD$logit.TEind.1ffF ~ kjønn + forskjell + kjønn:forskjell)

Coefficients:
         (Intercept)               kjønnJ         forskjellstor
kjønnJ:forskjellstor
            -0.25924              0.62987               0.05626
-0.61313


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.TEind.1ffF3)

                       Value p-value                Decision
```

```
Global Stat         2.628e+00  0.6219 Assumptions acceptable.
Skewness            6.197e-03  0.9373 Assumptions acceptable.
Kurtosis            1.195e+00  0.2743 Assumptions acceptable.
Link Function      -2.204e-15  1.0000 Assumptions acceptable.
Heteroscedasticity  1.427e+00  0.2323 Assumptions acceptable.
```

Anova-analyse av attributive adjektiver per substantiv (se side 338).

```
Call:
lm(formula = posD$AA.red.subst ~ forskjell)

Coefficients:
  (Intercept)  forskjellstor
    -0.002152       0.029169


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.AA.red.subst5)

                       Value p-value                Decision
Global Stat         1.821e+00  0.7686 Assumptions acceptable.
Skewness            1.539e+00  0.2148 Assumptions acceptable.
Kurtosis            9.510e-02  0.7578 Assumptions acceptable.
Link Function      -2.779e-16  1.0000 Assumptions acceptable.
Heteroscedasticity  1.873e-01  0.6652 Assumptions acceptable.
```

Anova-analyse av PC1 i prinsipalkomponentanalysen (s. 310).

```
Call:
lm(formula = pcaD.df$pc1 ~ forskjell)

Coefficients:
  (Intercept)  forskjellstor
       0.7548        -1.5095


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.pc1.8)

                       Value p-value                Decision
Global Stat         1.758e+00  0.7801 Assumptions acceptable.
Skewness            8.100e-01  0.3681 Assumptions acceptable.
Kurtosis            2.908e-02  0.8646 Assumptions acceptable.
Link Function      -2.707e-16  1.0000 Assumptions acceptable.
Heteroscedasticity  9.193e-01  0.3377 Assumptions acceptable.
```

Anova-analyse av PC2 i prinsipalkomponentanalysen (s. 310).

```
Call:
lm(formula = pcaD.df$pc2 ~ kjønn)

Coefficients:
(Intercept)       kjønnJ
```

```
    -0.4148       0.8296


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.pc2.3)

                        Value p-value                Decision
Global Stat         5.541e-01  0.9680 Assumptions acceptable.
Skewness            3.974e-01  0.5284 Assumptions acceptable.
Kurtosis            1.379e-01  0.7104 Assumptions acceptable.
Link Function      -1.801e-14  1.0000 Assumptions acceptable.
Heteroscedasticity  1.884e-02  0.8908 Assumptions acceptable.
```

Anova-analyse av PC3 i prinsipalkomponentanalysen (s. 310).

```
Call:
lm(formula = pcaD.df$pc3 ~ kjønn + lengde + forskjell + kjønn:forskjell)

Coefficients:
         (Intercept)               kjønnJ           lengdelang
              0.8860              -1.0346              -0.8014
       forskjellstor  kjønnJ:forskjellstor
             -0.5770               1.2821


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.pc3.4)

                       Value p-value                Decision
Global Stat         2.038968  0.7286 Assumptions acceptable.
Skewness            0.045227  0.8316 Assumptions acceptable.
Kurtosis            0.005552  0.9406 Assumptions acceptable.
Link Function       0.046596  0.8291 Assumptions acceptable.
Heteroscedasticity 1.941593  0.1635 Assumptions acceptable.
```

A5. Resultater fra Tukeys HSD-test

Anova-analyse av leksikalsk tetthet (s. 150).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.leksordF2)

$ferdighet
           diff         lwr        upr     p adj
S-M 0.005561277 -0.01506996 0.02619252 0.5913478

$forskjell
                 diff         lwr         upr     p adj
stor-liten -0.0187425 -0.03937374 0.001888738 0.0741283
```

```
$`ferdighet:forskjell`
                         diff          lwr          upr     p adj
S:liten-M:liten  0.029823103 -0.008743083  0.068389289 0.1832108
M:stor-M:liten   0.005519324 -0.033046862  0.044085510 0.9812790
S:stor-M:liten  -0.013181224 -0.051747410  0.025384961 0.8022235
M:stor-S:liten  -0.024303779 -0.062869965  0.014262407 0.3496957
S:stor-S:liten  -0.043004327 -0.081570513 -0.004438141 0.0231384
S:stor-M:stor   -0.018700549 -0.057266735  0.019865637 0.5768577
```

Anova-analyse av hapax legomena (s. 228).

```
> TukeyHSD(aov(lm.hapax4))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.hapax4)

$`kjønn`
           diff         lwr          upr     p adj
J-G -0.03150765 -0.06048804 -0.002527248 0.0336386

$forskjell
                  diff        lwr         upr     p adj
stor-liten -0.02171641 -0.0506968 0.007263992 0.1389411

$`kjønn:forskjell`
                          diff         lwr           upr     p adj
J:liten-G:liten -0.0636569546 -0.11783031 -0.0094836022 0.0150884
G:stor-G:liten  -0.0538657146 -0.10803907  0.0003076378 0.0518442
J:stor-G:liten  -0.0532240521 -0.10739740  0.0009493003 0.0558807
G:stor-J:liten   0.0097912400 -0.04438211  0.0639645924 0.9635194
J:stor-J:liten   0.0104329025 -0.04374045  0.0646062549 0.9563912
J:stor-G:stor    0.0006416625 -0.05353169  0.0548150149 0.9999888
```

Anova-analyse av MOS-entropi$_{W=100}$ (s. 337).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.MOSentropi100.2)

$lengde
                  diff          lwr         upr     p adj
lang-kort 0.0006097494 -0.002248137 0.003467635 0.6707234

$forskjell
                   diff          lwr         upr     p adj
stor-liten 0.0004461596 -0.002411726 0.003304045 0.7556426

$`lengde:forskjell`
                               diff           lwr         upr     p adj
lang:liten-kort:liten  4.636117e-03 -0.0008163195 0.010088554 0.1221274
kort:stor-kort:liten   4.584067e-03 -0.0008683694 0.010036504 0.1286929
lang:stor-kort:liten   9.815490e-04 -0.0038952586 0.005858357 0.9506778
kort:stor-lang:liten  -5.204992e-05 -0.0060248950 0.005920795 0.9999955
lang:stor-lang:liten  -3.654568e-03 -0.0091070048 0.001797868 0.2961245
lang:stor-kort:stor   -3.602518e-03 -0.0090549549 0.001849918 0.3084175
```

Anova-analyse av gjennomsnittlig antall leksikalske ord per klausus (s. 338).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.klLleks3)

$ferdighet
          diff        lwr       upr     p adj
S-M 0.06563608 -0.1799273 0.3111995 0.5944646

$forskjell
                 diff        lwr        upr     p adj
stor-liten -0.2163464 -0.4619098 0.02921694 0.0830339

$`ferdighet:forskjell`
                        diff        lwr         upr     p adj
S:liten-M:liten  0.32246337 -0.1365707  0.78149747 0.2568895
M:stor-M:liten   0.04048085 -0.4185533  0.49951495 0.9954644
S:stor-M:liten  -0.15071036 -0.6097445  0.30832375 0.8205953
M:stor-S:liten  -0.28198252 -0.7410166  0.17705158 0.3722949
S:stor-S:liten  -0.47317372 -0.9322078 -0.01413962 0.0409568
S:stor-M:stor   -0.19119120 -0.6502253  0.26784290 0.6892677
```

Anova-analyse av frekvens av korte subklaususer (s. 264).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.subklL3F.2)

$`kjønn`
             diff          lwr         upr     p adj
J-G -0.0006560543 -0.003527029 0.002214921 0.6487927

$lengde
                  diff          lwr           upr     p adj
lang-kort -0.003669154 -0.006540129 -0.0007981789 0.0132039

$forskjell
                   diff          lwr         upr    p adj
stor-liten -0.001048352 -0.003919326 0.001822623 0.467405

$`kjønn:forskjell`
                           diff          lwr         upr     p adj
J:liten-G:liten -0.0037283018 -0.009095862 0.001639258 0.2659057
G:stor-G:liten  -0.0041205991 -0.009488159 0.001246961 0.1882434
J:stor-G:liten  -0.0017044060 -0.007071966 0.003663154 0.8345344
G:stor-J:liten  -0.0003922973 -0.005759857 0.004975263 0.9973959
J:stor-J:liten   0.0020238958 -0.003343664 0.007391456 0.7506013
J:stor-G:stor    0.0024161931 -0.002951367 0.007783753 0.6340692
```

Anova-analyse av andel korte subklaususer (s. 265).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.subklL3subkl.3)

$ferdighet
          diff        lwr       upr     p adj
S-M -0.3411052 -0.8758527 0.1936423 0.2065815
```

```
$forskjell
                diff       lwr        upr     p adj
stor-liten -0.4952548 -1.030002 0.03949264 0.0688203

$`ferdighet:forskjell`
                        diff        lwr         upr     p adj
S:liten-M:liten  0.20346682 -0.7961421  1.20307569 0.9491107
M:stor-M:liten   0.04931716 -0.9502917  1.04892604 0.9991940
S:stor-M:liten  -0.83636002 -1.8359689  0.16324886 0.1315610
M:stor-S:liten  -0.15414966 -1.1537585  0.84545922 0.9767876
S:stor-S:liten  -1.03982684 -2.0394357 -0.04021796 0.0384910
S:stor-M:stor   -0.88567718 -1.8852861  0.11393170 0.0998804
```

Anova-analyse av andel t-enheter med kort forfelt (s. 292).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.TEind.1ffF3)

$`kjønn`
         diff       lwr       upr     p adj
J-G 0.3233078 0.0433251 0.6032906 0.0244069

$forskjell
                diff        lwr        upr    p adj
stor-liten -0.2503028 -0.5302855 0.02967996 0.078717

$`kjønn:forskjell`
                        diff        lwr         upr     p adj
J:liten-G:liten  0.62987487  0.1065003  1.15324946 0.0122499
G:stor-G:liten   0.05626423 -0.4671104  0.57963882 0.9918686
J:stor-G:liten   0.07300505 -0.4503695  0.59637965 0.9826170
G:stor-J:liten  -0.57361064 -1.0969852 -0.05023605 0.0264166
J:stor-J:liten  -0.55686982 -1.0802444 -0.03349522 0.0328484
J:stor-G:stor    0.01674083 -0.5066338  0.54011542 0.9997795
```

Anova-analyse av PC3 i prinsipalkomponentanalysen (s. 311).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm.pc3.4)

$`kjønn`
          diff        lwr        upr     p adj
J-G -0.4469707 -0.9741485 0.08020706 0.0949429

$lengde
                diff       lwr        upr     p adj
lang-kort -0.7423085 -1.269486 -0.2151307 0.0066316

$forskjell
                diff        lwr       upr     p adj
stor-liten 0.05291268 -0.4742651 0.5800905 0.8413269

$`kjønn:forskjell`
                        diff        lwr         upr     p adj
J:liten-G:liten -1.0843173 -2.0699262 -0.09870847 0.0256913
G:stor-G:liten  -0.5844339 -1.5700428  0.40117491 0.4033843
J:stor-G:liten  -0.3940580 -1.3796669  0.59155082 0.7154719
```

```
G:stor-J:liten   0.4998834 -0.4857255  1.48549224 0.5395843
J:stor-J:liten   0.6902593 -0.2953496  1.67586814 0.2591398
J:stor-G:stor    0.1903759 -0.7952329  1.17598475 0.9559529
```

A6. Prinsipalkomponenter

elev	pc1	pc2	pc3
202	2,0938912	-1,3931194	-0,6877332
203	-0,6277605	0,8432632	-1,4644888
205	3,9084888	-1,8037966	-0,7066891
206	-1,8372349	-0,8528219	1,7543082
208	1,5946973	-2,1571463	-1,1180294
210	-1,9202424	-0,19903	-1,155598
211	-0,5287306	-0,158568	3,1792298
212	-0,4371647	0,1252259	1,3583923
213	-0,9927977	-1,2654617	-0,4066866
215	0,4598227	0,5149906	1,569183
218	-0,1469832	0,6662732	0,5014392
222	-1,1394534	1,9123057	0,5952887
225	-1,6176971	0,1245692	1,6174346
226	-0,0140356	-0,2812099	-0,4489322
227	-1,2406066	-0,3588467	0,1612068
232	1,121843	-0,4168355	0,69793
233	0,2192301	3,8686822	-0,4550374
234	-2,2617739	2,908961	-0,0329809
235	0,8137748	1,5391113	2,0978151
236	0,5911294	0,3402187	0,4501574
237	-0,6302381	-3,0077706	0,2928992
238	-1,0983998	-0,8571699	1,0488282
239	-2,2086012	0,6555449	-0,1743968
241	-0,4075345	-1,5656759	-1,1894081
242	0,617502	-0,9956097	-0,7024271
243	4,4048162	1,9770818	0,0444693
245	1,4254129	0,2510714	-0,4661547
247	-3,2240497	-0,4663067	0,1512586
248	2,8559043	0,4649485	-0,5790308
249	-2,7302909	1,2037313	-0,9141861
251	1,6836727	-1,4453754	0,9108393
252	-0,0111377	0,6425748	1,309706
259	0,4589188	2,5537817	-1,8280932
260	-1,4350884	0,2316713	-0,1388698
261	-0,8998299	-2,1296971	1,0141918
264	-0,8701363	0,0623839	-3,0256734
265	-2,5507296	0,5731205	0,4076763
266	-2,152621	0,0805091	0,164817
269	-0,4913637	3,5259425	0,068195
272	-3,0001617	0,8779887	-2,7280583
273	-0,7016701	-1,8806642	0,0705561
278	1,2787024	1,5364076	0,9079577
281	1,266859	0,1712782	-1,1049945
285	2,071361	0,7748782	0,0647772
292	-2,1792129	-1,5312625	-0,9590382
293	-0,4083038	0,2539446	-1,460558
294	2,2346888	1,0837154	0,1244176
296	1,265142	-0,5066338	0,2231177
297	-0,8761743	-2,4167024	0,8791162
298	3,7130932	-0,6237869	-1,2733697
301	-0,3151184	-1,209566	-0,4281182
302	2,5258545	0,5134272	0,2239125
303	-0,7223493	-1,9835027	0,8651956

```
305  2,726332   1,2820381  0,9651607
307 -1,0735244  2,2440107  1,3141904
309  1,2224474 -2,5125811  0,9371864
310  1,0146667  0,1051665 -0,2045965
312  2,481215  -0,7352807 -1,0283933
313 -1,1318325  1,0884785 -0,3488757
317 -2,1666172 -2,2428749 -0,9404357
```

B. Prosessdokumenter

B1. Svar fra NSD

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

Harald Hårfagres gate 29
N-5007 Bergen
Norway
Tel: +47-55 58 21 17
Fax: +47-55 58 96 50
nsd@nsd.uib.no
www.nsd.uib.no
Org.nr. 985 321 884

Bård Jensen
Institutt for humanistiske fag
Avdeling for lærerutdanning og naturfag
Høgskolen i Hedmark
Postboks 4010 Bedriftssenteret
2306 HAMAR

Vår dato: 17.08.2007 Vår ref :17145/SM Deres dato: Deres ref:

KVITTERING PÅ MELDING OM BEHANDLING AV PERSONOPPLYSNINGER

Vi viser til melding om behandling av personopplysninger, mottatt 25.06.2007. Meldingen gjelder prosjektet:

17145	*Syntaktiske kjennetegn ved PC-skrevne og håndskrevne elevtekster*
Behandlingsansvarlig	*Høgskolen i Hedmark, ved institusjonens øverste leder*
Daglig ansvarlig	*Bård Jensen*

Personvernombudet har vurdert prosjektet og finner at behandlingen av personopplysninger er meldepliktig i henhold til personopplysningsloven § 31. Behandlingen tilfredsstiller kravene i personopplysningsloven.

Personvernombudets vurdering forutsetter at prosjektet gjennomføres i tråd med opplysningene gitt i meldeskjemaet, korrespondanse med ombudet, eventuelle kommentarer samt personopplysningsloven/-helseregisterloven med forskrifter. Behandlingen av personopplysninger kan settes i gang.

Det gjøres oppmerksom på at det skal gis ny melding dersom behandlingen endres i forhold til de opplysninger som ligger til grunn for personvernombudets vurdering. Endringsmeldinger gis via et eget skjema, http://www.nsd.uib.no/personvern/endringsskjema. Det skal også gis melding etter tre år dersom prosjektet fortsatt pågår. Meldinger skal skje skriftlig til ombudet.

Personvernombudet har lagt ut opplysninger om prosjektet i en offentlig database, http://www.nsd.uib.no/personvern/register/

Personvernombudet vil ved prosjektets avslutning, 31.12.2010, rette en henvendelse angående status for behandlingen av personopplysninger.

Vennlig hilsen

Bjørn Henrichsen Siv Midthassel

Kontaktperson: Siv Midthassel tlf: 55 58 83 34

Vedlegg: Prosjektvurdering

Personvernombudet for forskning

Prosjektvurdering - Kommentar

17145

Ombudet legger til grunn at prosjektleder ikke får utlevert elevtekster som inneholder sensitive opplysninger eller opplysninger om tredjeperson, som avtalt med prosjektleder på telefon 17.08.2007.

Datamaterialet anonymiseres ved prosjektslutt ved at verken direkte eller indirekte personidentifiserbare opplysninger fremgår, navneliste og manuelle tekster slettes og evt indirekte personidentifiserbare variabler fjernes/tilstrekkelig kategoriseres. Prosjektslutt er satt til 31.12.2010.

Det legges til grunn at elevtekster som arkiveres elektronisk for evt videre forskning foreligger i anonymisert form og ikke kan tilbakeføres til enkeltpersoner.

B2. Rettledning for lærere

Rettledning for lærer

Deltagelse i forskningsprosjekt om pc-skrevne elevtekster

Til: Norsklærer ved <skole> videregående skole

Fra: Bård Uri Jensen, stipendiat ved Høgskolen i Hedmark

Dato: 2009-01-21

Dette notatet er en oppskrift til lærere som bidrar med datainnsamling til prosjektet "Språklige forhold i pc-skrevne elevtekster". Det gir en oversikt over de nødvendige betingelser for gjennomføring i klassen. Mer informasjon om prosjektet finnes i et eget notat. Ta gjerne kontakt med meg for en nærmere orientering eller hvis det er noe du lurer på.

Kontakt: epost: Bard.Jensen@hihm.no / telefon arbeid: 625 17826 / mobil: 482 68986 / fax: 625 17601
Postadresse: Høgskolen i Hedmark, Lærerskolealleen 1, 2418 Elverum.

FØR OPPSTART

1. Vi har en samtale om prosjektet og den praktiske gjennomføringen.
2. Jeg sender deg riktig antall orienteringsbrev med svarslipp (2 ark per elev), samt spørreskjema (2 ark per elev). Svarslippene og spørreskjemaene er ferdig påført et unikt id-nummer for hver elev. Med i forsendelsen er også et forsendelsesskjema.

Informasjonsskriv og samtykke

Elever som skal delta i prosjektet, må gi sitt samtykke:

3. Del ut brevet med svarslipp til alle aktuelle elever, gjerne supplert med en kort muntlig orientering.
4. Samle inn - og eventuelt purr - svarslippen. Foreldrene trenger ikke signere.
5. Bruk klasselisten din og noter det ferdigutfylte id-nummeret fra svarslippen fra hver enkelt elev som ønsker å delta. (Denne listen skal ikke jeg se.)
6. Bruk den navnløse klasselisten du får av meg, til å krysse av hvilke elever som samtykker, inkludert om de samtykker i senere bruk.

Spørreskjema

Når alle svarslippene er kommet inn, må det settes av litt tid til å fylle ut spørreskjemaet. Skjemaet er enkelt og bør ikke ta mer enn 10 minutter.

7. Del ut spørreskjema med riktig id-nummer til hver elev. (Nummer er ferdig påført av meg, men du må finne riktig elev til hvert nummer på klasselisten din.)
8. Gi de utfylte skjemaene og samtykkelistene til <fagkoordinator>, så henter jeg dem hos henne.

SKRIVEØKTENE

1. Det skal være to skriveøkter på to ulike dager, gjerne med noen dagers eller ukers mellomrom.
2. En skriveøkt bør være på ca. 2 skoletimer.

3. Oppgave A1 brukes den første dagen, og oppgave A2 den andre dagen. Oppgavetekstene får dere av meg.
4. Den første dagen skriver den ene halvparten av elevene for hånd og den andre med pc. Den andre dagen bytter de på, slik at den andre halvparten skriver for hånd og den første med pc. Dersom flere klasser er involvert, kan denne inndelingen gjerne gjøres klassevis.
5. Elevene kan arbeide prosessorientert og bruke hverandre som responsgivere. Ellers arbeider dere slik dere pleier i skriveøktene.
6. Dersom det er praktisk mulig, bør elevene ikke ha tilgang til Internett under skriveøktene.
7. Elevene merker teksten sin med sitt id-nummer (fra svarslippen/spørreskjema) i stedet for navn. (Id-nummeret har du på klasselisten.)

ETTER SKRIVINGEN

1. Dersom noen elever har skrevet navnet sitt på teksten, må du stryke ut dette og erstatte det med id-nummeret fra klasselisten.
2. Kopier alle håndskrevne tekster på kopimaskin. (Hvis du enkelt kan skanne tekstene til pdf, er det enda greiere.) Dette bør du gjøre *før* du retter, slik at jeg ikke får se dine rettelser eller kommentarer.
3. Hvis noen av tekstene inneholder sensitive opplysninger om eleven eller om tredje-person, skal de av personvernhensyn ikke sendes til meg. Du må altså lese/rette tekstene *før* du sender dem til meg.
4. Fyll ut det vedlagte forsendelsesskjemaet.
5. Send elektroniske kopier av pc-tekstene på e-post. Send håndtekstene som pdf på e-post eller gi papirkopier til <fagkoordinator>. Legg ved det utfylte forsendelsesskjemaet.

HJELP TIL GJENNOMFØRINGEN

Skriveøktene er ment å skulle inngå i det ordinære undervisningsarbeidet i skolehverdagen. Kopiering og annet praktisk tilleggsarbeid kan jeg komme til skolen og utføre dersom du ønsker det.

B3. Orienteringsbrev til elevene

Høgskolen i Hedmark

Telefon: 625 17826
E-post: bard.jensen@hihm.no

Til elever i VG1 ved xxx videregående skole

Hamar, 21. januar 2009

Deltakelse i forskningsprosjektet "Språklige forhold i pc-skrevne elevtekster"

Forskningsprosjektet "Språklige forhold i pc-skrevne elevtekster", som jeg gjennomfører ved Høgskolen i Hedmark i perioden 1.8.2006-31.01.2011, har som formål å undersøke *språket* i skolearbeid som elever har gjort på pc. I den forbindelse ber jeg om å få bruke tekster som dere skal skrive i norskfaget dette skoleåret. For at jeg skal kunne gjøre de analysene jeg ønsker, er det nødvendig at dere skriver noen av tekstene for hånd. Elever som deltar i prosjektet, vil også måtte svare på et kort spørreskjema.

Navnet ditt vil ikke stå på tekstene, og bare norsklæreren vil vite hvilke tekster du har skrevet. Jeg vil ikke kunne koble verken tekster eller forskningsresultater til enkeltelever. Jeg vil ikke være til stede når du skriver, og jeg vil ikke kjenne identiteten din. Når jeg skal analysere tekstene, vil jeg skrive av det du har skrevet for hånd, og lagre det elektronisk. Deretter vil jeg blant annet ved hjelp av datamaskin undersøke *språklige forhold* i alle tekstene. Jeg skal altså ikke analysere hva du har skrevet om, men bare språket.

Bare jeg og mine to veiledere vil ha tilgang til dataene og tekstene. Vi har taushetsplikt, og tekstene og annen informasjon som blir samlet inn gjennom prosjektet, blir behandlet konfidensielt. Dessuten vil læreren din lese alle tekstene dine før jeg får dem, og ikke gi meg tekster som inneholder sensitive opplysninger om deg selv eller andre. Prosjektet er godkjent av Personvernombudet for forskning, Norsk samfunnsvitenskapelig datatjeneste AS.

Når prosjektet avsluttes, vil alle data bli fullstendig anonymisert. Det vil si at jeg makulerer de håndskrevne arkene, og i tillegg at læreren din makulerer listene der tekstene dine er koblet til navnet ditt. Etter dette vil verken norsklæreren eller andre kunne finne ut hvem som har skrevet hvilke tekster. Tekstene vil jeg imidlertid arkivere elektronisk for at flere språklige forhold skal kunne undersøkes i eventuelle senere prosjekter.

Forskning av denne typen krever tillatelse fra eleven. Jeg ber deg derfor om å fylle ut og returnere svarslippen til norsklæreren din snarest og senest 2. februar. Det er frivillig å delta, og du kan når som helst trekke tillatelsen tilbake ved å kontakte meg. Hvorvidt du ønsker å delta eller ikke, vil ikke ha noen innvirkning på undervisningen eller ditt forhold til skolen på andre måter.

I tillegg til å gi meg tillatelse til å bruke tekstene dine i dette prosjektet, kan du på svarslippen også gi tillatelse til at tekstene og opplysninger fra spørreskjema kan brukes i andre prosjekter også av andre forskere. Det er frivillig å gi slik tillatelse, og du kan godt være med i prosjektet mitt uten å gi slik tillatelse. Dersom du ikke krysser av her, vil ingen andre forskere få anledning til å se tekstene dine.

Med vennlig hilsen

Bård Uri Jensen
Doktorgradsstipendiat

Avd. for lærerutdanning og naturfag
Institutt for humanistiske fag

Postadresse:
Postboks 4010 Bedr, 2306 Hamar

Telefon:
62 51 76 00

Telefaks:
62 51 76 01

B4. Skjema for samtykke

Høgskolen i Hedmark

Telefon: 625 17826
E-post: bard.jensen@hihm.no

Deltakelse i forskningsprosjektet "Språklige forhold i pc-skrevne elevtekster"

Elev nr «id2»

Svarslipp

Samtykkeerklæring

Jeg, (navn) _____, har fått en skriftlig orientering om forskningsprosjektet "Språklige forhold i pc-skrevne elevtekster" og gir tillatelse til at tekster jeg skriver i forbindelse med norskfaget dette skoleåret, blir benyttet i prosjektet.

kryss av

☐ Jeg gir også tillatelse til at mine anonymiserte tekster og anonymiserte data om meg fra spørreskjema kan brukes i senere prosjekter også av andre forskere.

_____ _____
Dato Signatur

Jeg, (navn) _____, ønsker ikke at mine tekster benyttes i forskningsprosjektet "Språklige forhold i pc-skrevne elevtekster".

_____ _____
Dato Signatur

Svarslippen leveres til norsklæreren senest 2. februar.

B5. Forsendelsesskjema for lærere

Forsendelsesskjema elevtekster

Språklige forhold i pc-skrevne elevtekster *<skole>*

Sendes til Bård Uri Jensen, Høgskolen i Hedmark.

Postadresse: Postboks 4010 Bedriftssenteret, 2306 Hamar.
E-post: Bard.Jensen@hihm.no / telefon arbeid: 625 17826 / mobil: 482 68986.

Oppgave: A1 ○ A2 ○ B1 ○ B2 ○
C1 ○ C2 ○ D1 ○ D2 ○

Oppgavetekst: _____

Tidsramme: 2 skoletimer ○ 7 dager ○

Eventuelt avvik eller merknad:

_____

_____

Hvis teksten er skrevet på 7 dager:

Skrevet på skolen ○ Skrevet hjemme ○ Skrevet på skolen og hjemme ○

Merknad:

_____

_____

Tilgang til Internett under skrivingen:

Ja ○ Nei ○

Merknad:

_____

_____

Tekster innlevert dato: ____.____._____

Lærer/klasse: _____

B6. Spørreskjema A

Side 1 av 4 Elev nr «id2»

Spørreskjema

"Språklige forhold i pc-skrevne elevtekster"

Dine holdninger til pc

1. Hvor godt synes du at du behersker tekstbehandlingsprogrammet (Word e.l.) på pc-en?

bra ○ ganske bra ○ ganske dårlig ○ dårlig ○

2. Når du arbeider hjemme med norskfaget, foretrekker du å skrive på pc eller for hånd?

alltid pc ○ oftest pc ○ oftest hånd ○ alltid hånd ○

Hvis du svarte "alltid pc" eller "oftest pc" på spørsmål 2, svarer du på spørsmålene i 3a.

3a. Hvorfor foretrekker du å skrive på pc i forbindelse med norskfaget?

	viktig	litt viktig	ikke viktig	usant
Det går raskere:	○	○	○	○
Det blir penere:	○	○	○	○
Jeg tenker bedre:	○	○	○	○
Det er lettere å gjøre endringer:	○	○	○	○
Det er lettere å lage tegninger og figurer:	○	○	○	○
Pga. stavekontrollen:	○	○	○	○
Det er lettere å kopiere fra Internett:	○	○	○	○
Det er lettere å samarbeide med andre elever:	○	○	○	○

Bård Uri Jensen, Høgskolen i Hedmark. bard.jensen@hihm.no, 625 17826

Hvis du svarte "alltid hånd" eller " oftest hånd" på spørsmål 2, svarer du på spørsmålene i 3b.

3b. Hvorfor foretrekker du å skrive for hånd i forbindelse med norskfaget?

	viktig	litt viktig	ikke viktig	usant
Det går raskere:	○	○	○	○
Det blir penere:	○	○	○	○
Jeg tenker bedre:	○	○	○	○
Det er lettere å gjøre endringer:	○	○	○	○
Det er lettere å lage tegninger og figurer:	○	○	○	○
Det er enklere å lage tankekart, disposisjon, etc.:	○	○	○	○
Jeg har ikke pc hjemme:	○	○	○	○

Din bruk av pc

4. Hva bruker du pc-en til hjemme? (*markér ett alternativ for hver aktivitet*)

	mye	en del	lite	aldri
Skriftlig skolearbeid:	○	○	○	○
E-post, *MSN*, *Facebook*, annen skriftlig kommunikasjon:	○	○	○	○
Annen skriving:	○	○	○	○
Skype eller annen muntlig kommunikasjon:	○	○	○	○
Surfing på *www*:	○	○	○	○
Spill:	○	○	○	○

Hvilke spill?: _____

Annet:	○	○	○	○

Hva da?: _____

5. Hvor ofte bruker du pc hjemme til skolearbeid eller fritid? (*markér ett alternativ*)

- ○ Hver dag eller nesten hver dag
- ○ 1-4 dager i uka
- ○ Sjeldnere enn 1 dag i uka

6. Hvis du tenker bare på de dagene du bruker pc hjemme, hvor lenge sitter du da ved pc-en en typisk dag? (*markér ett alternativ*)

- ○ 0-30 minutter
- ○ ½ - 2 timer
- ○ Over 2 timer

7. Når du bruker Word (eller det tekstbehandlingsprogrammet som du pleier å bruke), hvordan synes du det er å...

	lett	litt vanskelig
skrive inn tekst:	○	○
rette feiltastinger:	○	○
sette inn et nytt ord i en setning:	○	○
sette inn en ny setning:	○	○
sette inn et nytt avsnitt:	○	○
slette et ord i en setning:	○	○
slette en setning:	○	○
slette et avsnitt:	○	○
flytte et ord i en setning:	○	○
flytte en setning:	○	○
flytte et avsnitt:	○	○
bruke angre-funksjonen:	○	○

Personlige opplysninger

8. Kjønn
○ Kvinne
○ Mann

9. Morsmål: (*markér ett eller flere alternativ*)
☐ Norsk
☐ Annet

10. Målform som hovedmål: (*markér ett alternativ*)
○ Bokmål
○ Nynorsk

11. *Standpunktkarakter* i norsk hovedmål på avgangsvitnemålet fra grunnskolen:
6:○ 5:○ 4:○ 3:○ 2:○ 1:○

Bård Uri Jensen, Høgskolen i Hedmark. bard.jensen@hihm.no, 625 17826

B7. Spørreskjema B

Side 1 av 2 Elev nr «id2»

Spørreskjema B

"Språklige forhold i pc-skrevne elevtekster"

1. Hvor godt liker du å skrive <u>for hånd</u> når du skriver i forbindelse med norskfaget?

godt	ganske godt	ganske dårlig	dårlig
○	○	○	○

2. Her er noen påstander om deg når du skriver for hånd. Hvor enig eller uenig er du i hver av påstandene?

(Markér ett svaralternativ for hver påstand.)

	enig	litt enig	litt uenig	uenig
Det går greit.	○	○	○	○
Jeg synes det er slitsomt å bruke penn eller blyant.	○	○	○	○
Jeg skriver så lite som mulig.	○	○	○	○
Jeg skriver raskt.	○	○	○	○
Jeg får konsentrert meg godt om det jeg skriver om.	○	○	○	○
Jeg strever med å skrive pent nok.	○	○	○	○
Jeg er ofte usikker på hvordan jeg skal stave ordene.	○	○	○	○
Jeg forandrer mye, slik at teksten skal bli best mulig.	○	○	○	○
Når jeg skriver tekster i norskfaget, skriver jeg stikkord, tegner tankekart, e.l. før jeg begynner selve skrivingen.	○	○	○	○

3. Hvor godt synes du at du greier å lage tekster når du skriver dem for hånd?

bra	ganske bra	ganske dårlig	dårlig
○	○	○	○

Bård Uri Jensen, Høgskolen i Hedmark. bard.jensen@hihm.no, 625 17826

4. Hvor ofte skriver du noe for hånd hjemme utenom skolearbeid?
(*Markér ett alternativ.*)

○ Hver dag eller nesten hver dag
○ 1-4 dager i uka
○ Sjeldnere enn 1 dag i uka

5. Hva skriver du for hånd utenom skolearbeid? (*F.eks.: dagbok, brev/kort, huskelister, beskjeder, fyller ut skjemaer, ...*

Ofte: _____

Av og til: _____

6. Hvis du legger sammen *alt* du skrev for hånd i går utenom skolearbeid, hvor mye ville det bli til sammen?
(*Markér ett alternativ.*)

○ Ingenting
○ Noen få ord
○ Noen setninger
○ 1/2 – 2 sider
○ Mer enn 2 sider

7. Hvis du legger sammen *alt* du skrev på pc i går utenom skolearbeid, hvor mye ville det bli til sammen?
(*Markér ett alternativ.*)

○ Ingenting
○ Noen få ord
○ Noen setninger
○ 1/2 – 2 sider
○ Mer enn 2 sider

8. Hvor lenge satt du ved pc-en hjemme i går?
(*Markér ett alternativ.*)

○ Ikke i det hele tatt
○ 0 – 30 minutter
○ 1/2 – 2 timer
○ Over 2 timer

9. Hvilket språk bruker du mest hjemme:
(*Markér ett eller flere alternativ.*)

☐ Norsk
☐ Annet: _____

Bård Uri Jensen, Høgskolen i Hedmark. bard.jensen@hihm.no, 625 17826

B8. Oppgave A1

OPPGAVE A1 "BØKER ELLER DATA?"

Tidsramme: 2 skoletimer
Sjanger: leserinnlegg
Målform: bokmål

I et leserinnlegg om ungdommers medievaner ble følgende påstand fremsatt: «Gutter leser ikke bøker, men driver med data. Jenter driver ikke med data, men leser bøker».

Skriv et leserinnlegg til en avis der du kommenterer og drøfter denne påstanden. Bruk overskriften "Bøker eller data?".

B9. Oppgave A2

OPPGAVE A2 "UNGDOMSFYLLA?"

Tidsramme: 2 skoletimer
Sjanger: leserinnlegg
Målform: bokmål

En MMI-undersøkelse viser at ungdommens drikkevaner har endret seg i negativ retning. I en lederartikkel i Hamar Arbeiderblad 6. november 2006 skriver redaktøren blant annet:

> *Hva gjør foreldrene når 14-15-åringene kommer og ber om øl til kveldens fest? Svaret bør være enkelt.*

Skriv et leserinnlegg der du først gjør greie for hva du tror kan være årsakene til den negative utviklingen, og deretter diskuterer synspunktet til redaktøren. Bruk overskriften "Ungdomsfylla?".

C. Korpus

C1. Document type definition (dtd)

Dtd-definisjonene som definerer de spesialtilpassede feltene i korpuset.

```
<!ELEMENT TEI.2 (#PCDATA | teiHeader | text | p)*>
<!ENTITY amp "&#38;">
    <!ATTLIST TEI.2 id CDATA #IMPLIED>
<!ELEMENT note (#PCDATA)>
    <!ATTLIST note type CDATA #IMPLIED>
    <!ATTLIST note resp CDATA #IMPLIED>
<!ELEMENT hi (#PCDATA | corr | add | del)*>
    <!ATTLIST hi rend CDATA #IMPLIED>
<!ELEMENT teiHeader (#PCDATA | fileDesc | profileDesc | revisionDesc)*>
<!ELEMENT profileDesc (textClass | particDesc)*>
<!ELEMENT textClass (catRef*)>
<!ELEMENT particDesc (person*)>
<!ELEMENT person EMPTY>
  <!ATTLIST person personid CDATA #IMPLIED >
  <!ATTLIST person standard (ja | nei) "ja" >
  <!ATTLIST person karakter ( 6 | 5 |4 | 3 | 2 | 1 | ukjent ) "ukjent" >
  <!ATTLIST person skriveferdighet ( høy | middels | lav ) "middels" >
  <!ATTLIST person kjoenn (kvinne | mann) "kvinne" >
  <!ATTLIST person trinn ( VG2 | VG1 | 10. ) "VG1" >
  <!ATTLIST person s1 ( norsk | annet | flere ) "norsk" >
  <!ATTLIST person hovedmaal (bokmål | nynorsk ) "bokmål" >
  <!ATTLIST person dialekt (ukjent) "ukjent" >
<!ELEMENT catRef EMPTY>
    <!ATTLIST catRef target CDATA #IMPLIED>
<!ELEMENT author (#PCDATA)>
<!ELEMENT corr (#PCDATA | corr | add | del | hi | unclear)*>
    <!ATTLIST corr sic CDATA #IMPLIED>
<!ELEMENT head (#PCDATA | corr | del | add)*>
    <!ATTLIST head type CDATA #IMPLIED>
<!ELEMENT front (#PCDATA | div)*>
<!ELEMENT del (#PCDATA | hi | add)*>
    <!ATTLIST del rend CDATA #IMPLIED>
<!ELEMENT body (#PCDATA | div)*>
<!ELEMENT change (#PCDATA | date | respStmt)*>
<!ELEMENT text (#PCDATA | front | body | back)*>
   <!ATTLIST text tekstid CDATA #IMPLIED >
   <!ATTLIST text oppgave ( A1 | A2 | B1 | B2 | C1 | C2 | D1 | D2 | A3 |
B3 | C3 | D3 ) "A1" >
   <!ATTLIST text oppgavetekst CDATA #IMPLIED >
   <!ATTLIST text tidsramme (kort | lang ) "kort" >
   <!ATTLIST text tid CDATA #IMPLIED >
   <!ATTLIST text sjanger (argumenterende | narrativ ) "argumenterende" >
   <!ATTLIST text modus (hånd | tastatur) "hånd" >
<!ELEMENT div (#PCDATA | head | p | lb )*>
    <!ATTLIST div type CDATA #IMPLIED>
<!ELEMENT add (#PCDATA | corr | hi | del)*>
    <!ATTLIST add type CDATA #IMPLIED>
    <!ATTLIST add resp CDATA #IMPLIED>
<!ELEMENT unclear (#PCDATA)>
    <!ATTLIST unclear reason CDATA #IMPLIED>
<!ELEMENT abbr (#PCDATA)>
<!ELEMENT resp (#PCDATA)>
```

```
<!ELEMENT p (#PCDATA | t-unit | clause | frag | s | lb | note | hi | abbr
| corr | del | unclear | add)*>
 <!ELEMENT s (#PCDATA | lb | note | hi | abbr | corr | del | unclear |
add)*>
<!ELEMENT TEMP (#PCDATA | t-unit | clause | frag | lb | note | hi | abbr |
corr | del | unclear | add)*>
<!ELEMENT t-unit (#PCDATA | TEMP | clause | frag | lb | note | hi | abbr |
corr | del | unclear | add)*>
 <!ATTLIST t-unit type ( imp | spm | frag ) #IMPLIED >
<!ELEMENT clause (#PCDATA | TEMP | clause | lb | note | hi | abbr | corr |
del | unclear | add)*>
    <!ATTLIST clause type ( relativ | nominal | adverbial | helsetning )
"relativ">
    <!ATTLIST clause subtype ( tid | sted | årsak | betingelse | hensikt |
følge | innrømmelse | sammenligning ) "tid">
<!ELEMENT frag (#PCDATA | TEMP | clause | lb | note | hi | abbr | corr |
del | unclear | add)*>
<!ELEMENT back (#PCDATA | div)*>
<!ELEMENT lb (#PCDATA)>
<!ELEMENT fileDesc (#PCDATA | titleStmt)*>
<!ELEMENT respStmt (#PCDATA | name | resp)*>
<!ELEMENT date (#PCDATA)>
<!ELEMENT titleStmt (#PCDATA | title | author )*>
<!ELEMENT revisionDesc (#PCDATA | change)*>
<!ELEMENT title (#PCDATA)>
<!ELEMENT name (#PCDATA)>
```

D. Elever og tekster

D1. Elev- og overordnet tekstinformasjon

Tekst	Elev	Oppgave	Modus	Kjønn	Ferdighet	Lengde	Forskjell	Karakter	Ord1	Ord3
A2-202	202	A2	H	J	M	lang	liten	4	471	471
A1-202	202	A1	T	J	M	lang	liten	4	557	554
A1-203	203	A1	H	J	M	kort	liten	4	431	426
A2-203	203	A2	T	J	M	kort	liten	4	442	442
A2-205	205	A2	H	J	S	kort	liten	5	410	409
A1-205	205	A1	T	J	S	kort	liten	5	457	450
A1-206	206	A1	H	G	S	lang	liten	5	512	508
A2-206	206	A2	T	G	S	lang	liten	5	526	520
A2-208	208	A2	H	G	S	lang	liten	5	661	659
A1-208	208	A1	T	G	S	lang	liten	5	714	714
A1-210	210	A1	H	G	S	lang	stor	5	648	638
A2-210	210	A2	T	G	S	lang	stor	5	1491	1483
A2-211	211	A2	H	G	M	kort	liten	4	330	326
A1-211	211	A1	T	G	M	kort	liten	4	218	216
A1-212	212	A1	H	J	M	kort	liten	4	449	445
A2-212	212	A2	T	J	M	kort	liten	4	511	508
A2-213	213	A2	H	J	M	lang	stor	4	407	406
A1-213	213	A1	T	J	M	lang	stor	4	724	722
A2-215	215	A2	H	G	M	kort	liten	3	296	293
A1-215	215	A1	T	G	M	kort	liten	3	256	253
A1-218	218	A1	H	G	S	lang	stor	5	454	453
A2-218	218	A2	T	G	S	lang	stor	5	800	793
A1-222	222	A1	H	G	M	lang	stor	4	597	595
A2-222	222	A2	T	G	M	lang	stor	4	757	753
A1-225	225	A1	H	J	M	kort	stor	4	242	238
A2-225	225	A2	T	J	M	kort	stor	4	467	463
A2-226	226	A2	H	G	M	kort	stor	4	267	266
A1-226	226	A1	T	G	M	kort	stor	4	359	359
A1-227	227	A1	H	G	S	lang	stor	6	388	387
A2-227	227	A2	T	G	S	lang	stor	6	1188	1182
A2-232	232	A2	H	G	M	kort	liten	4	316	315
A1-232	232	A1	T	G	M	kort	liten	4	361	357
A1-233	233	A1	H	J	S	lang	stor	5	669	664
A2-233	233	A2	T	J	S	lang	stor	5	1279	1261
A1-234	234	A1	H	J	S	lang	stor	5	245	245
A2-234	234	A2	T	J	S	lang	stor	5	877	870
A2-235	235	A2	H	J	M	kort	liten	4	351	344
A1-235	235	A1	T	J	M	kort	liten	4	344	342
A2-236	236	A2	H	G	S	kort	liten	5	434	432
A1-236	236	A1	T	G	S	kort	liten	5	476	476
A1-237	237	A1	H	G	M	lang	stor	4	445	425
A2-237	237	A2	T	G	M	lang	stor	4	561	560
A1-238	238	A1	H	G	M	kort	stor	4	280	280

A2-238	238	A2	T	G	M	kort	stor	4	601	597
A1-239	239	A1	H	G	S	kort	stor	5	275	274
A2-239	239	A2	T	G	S	kort	stor	5	657	654
A1-241	241	A1	H	G	S	kort	stor	5	291	291
A2-241	241	A2	T	G	S	kort	stor	5	502	502
A2-242	242	A2	H	G	M	kort	liten	4	305	305
A1-242	242	A1	T	G	M	kort	liten	4	378	374
A2-243	243	A2	H	J	S	kort	liten	5	459	458
A1-243	243	A1	T	J	S	kort	liten	5	463	460
A2-245	245	A2	H	J	S	lang	liten	5	516	514
A1-245	245	A1	T	J	S	lang	liten	5	518	516
A1-247	247	A1	H	G	M	kort	liten	4	371	369
A2-247	247	A2	T	G	M	kort	liten	4	445	442
A2-248	248	A2	H	J	M	kort	liten	4	330	329
A1-248	248	A1	T	J	M	kort	liten	4	376	371
A1-249	249	A1	H	J	M	lang	stor	4	352	349
A2-249	249	A2	T	J	M	lang	stor	4	756	756
A2-251	251	A2	H	G	M	kort	stor	4	252	252
A1-251	251	A1	T	G	M	kort	stor	4	406	405
A2-252	252	A2	H	G	S	kort	liten	6	362	359
A1-252	252	A1	T	G	S	kort	liten	6	421	412
A2-259	259	A2	H	G	M	lang	liten	4	525	519
A1-259	259	A1	T	G	M	lang	liten	4	462	460
A1-260	260	A1	H	J	S	lang	stor	5	514	508
A2-260	260	A2	T	J	S	lang	stor	5	848	840
A1-261	261	A1	H	J	S	lang	stor	6	391	386
A2-261	261	A2	T	J	S	lang	stor	6	815	803
A1-264	264	A1	H	J	S	lang	liten	5	548	547
A2-264	264	A2	T	J	S	lang	liten	5	640	637
A1-265	265	A1	H	G	M	kort	stor	3	276	273
A2-265	265	A2	T	G	M	kort	stor	3	544	534
A1-266	266	A1	H	G	S	lang	stor	6	472	469
A2-266	266	A2	T	G	S	lang	stor	6	599	596
A1-269	269	A1	H	J	S	kort	stor	5	315	315
A2-269	269	A2	T	J	S	kort	stor	5	641	634
A1-272	272	A1	H	G	S	lang	stor	5	358	355
A2-272	272	A2	T	G	S	lang	stor	5	622	609
A2-273	273	A2	H	G	S	lang	liten	6	509	504
A1-273	273	A1	T	G	S	lang	liten	6	542	536
A2-278	278	A2	H	G	M	kort	liten	4	361	359
A1-278	278	A1	T	G	M	kort	liten	4	358	354
A2-281	281	A2	H	J	S	lang	liten	5	597	592
A1-281	281	A1	T	J	S	lang	liten	5	688	684
A2-285	285	A2	H	J	M	lang	stor	4	644	641
A1-285	285	A1	T	J	M	lang	stor	4	857	849
A1-292	292	A1	H	J	M	kort	liten	4	423	421
A2-292	292	A2	T	J	M	kort	liten	4	523	522
A1-293	293	A1	H	J	M	lang	liten	4	572	559

A2-293	293	A2	T	J	M	lang	liten	4	545	541
A2-294	294	A2	H	J	M	kort	stor	4	371	371
A1-294	294	A1	T	J	M	kort	stor	4	465	464
A2-296	296	A2	H	G	M	lang	liten	4	501	497
A1-296	296	A1	T	G	M	lang	liten	4	498	495
A2-297	297	A2	H	G	S	lang	liten	5	488	484
A1-297	297	A1	T	G	S	lang	liten	5	512	506
A2-298	298	A2	H	J	S	kort	liten	5	349	347
A1-298	298	A1	T	J	S	kort	liten	5	185	184
A1-301	301	A1	H	J	M	lang	stor	4	504	501
A2-301	301	A2	T	J	M	lang	stor	4	680	668
A2-302	302	A2	H	J	S	kort	liten	5	389	386
A1-302	302	A1	T	J	S	kort	liten	5	480	478
A1-303	303	A1	H	G	M	kort	stor	4	295	295
A2-303	303	A2	T	G	M	kort	stor	4	406	404
A2-305	305	A2	H	J	S	lang	liten	5	534	534
A1-305	305	A1	T	J	S	lang	liten	5	474	473
A1-307	307	A1	H	J	M	kort	stor	4	196	195
A2-307	307	A2	T	J	M	kort	stor	4	399	397
A2-309	309	A2	H	G	S	kort	liten	5	464	462
A1-309	309	A1	T	G	S	kort	liten	5	433	432
A2-310	310	A2	H	J	S	kort	stor	5	240	237
A1-310	310	A1	T	J	S	kort	stor	5	385	381
A2-312	312	A2	H	G	S	lang	stor	5	418	415
A1-312	312	A1	T	G	S	lang	stor	5	1157	1135
A1-313	313	A1	H	J	M	lang	stor	4	324	324
A2-313	313	A2	T	J	M	lang	stor	4	949	946
A1-317	317	A1	H	J	S	lang	stor	5	481	455
A2-317	317	A2	T	J	S	lang	stor	5	762	749

D2. Spørreskjemasvar

Grunnleggende elevegenskaper.

ID	Kjønn	Morsmål	Karakter	Word	Foretrekker	Hvor_ofte	Hvor_lenge
202	J	Norsk	4	Bra	Oftest pc	Hver dag	1/2 - 2 t
203	J	Norsk	4	Ganske bra	Oftest pc	1-4 dager	over 2 t
205	J	Norsk	5	Ganske bra	Oftest hånd	Hver dag	1/2 - 2 t
206	G	Norsk	5	Ganske bra	Oftest pc	Hver dag	1/2 - 2 t
208	G	Norsk	5	Ganske bra	Alltid pc	Hver dag	1/2 - 2 t
210	G	Norsk	5	Bra	Alltid pc	Hver dag	1/2 - 2 t
211	G	Norsk	4	Bra	Oftest pc	Hver dag	over 2 t
212	J	Norsk	4	Ganske bra	Alltid pc	Hver dag	1/2 - 2 t
213	J	Norsk	4	Ganske bra	Oftest pc	Hver dag	1/2 - 2 t
215	G	Norsk	3	Ganske bra	Alltid pc	Hver dag	1/2 - 2 t
218	G	Norsk	5	Bra	Alltid pc	Hver dag	1/2 - 2 t
222	G	Norsk	4	Bra	Alltid pc	Hver dag	over 2 t

225	J	Norsk	4	Ganske bra	Oftest pc	Hver dag	1/2 - 2 t
226	G	Norsk	4	Ganske bra	Oftest pc	1-4 dager	1/2 - 2 t
227	G	Norsk	6	Bra	Oftest pc	Hver dag	1/2 - 2 t
232	G	Norsk	4	Bra	Oftest pc	Hver dag	over 2 t
233	J	Norsk	5	Bra	Oftest pc	Hver dag	1/2 - 2 t
234	J	Norsk	5	Bra	Oftest pc	Hver dag	1/2 - 2 t
235	J	Norsk	4	Ganske bra	Oftest pc	Hver dag	over 2 t
236	G	Norsk	5	Ganske bra	Alltid pc	Hver dag	over 2 t
237	G	Norsk	4	Bra	Alltid pc	Hver dag	over 2 t
238	G	Norsk	4	Bra	Alltid pc	Hver dag	over 2 t
239	G	Norsk	5	Bra	Alltid pc	Hver dag	1/2 - 2 t
241	G	Norsk	5	Bra	Alltid pc	Hver dag	1/2 - 2 t
242	G	Norsk	4	Bra	Oftest hånd	Hver dag	over 2 t
243	J	Norsk	5	Ganske bra	Oftest pc	Hver dag	over 2 t
245	J	Norsk	5	Bra	Oftest pc	Hver dag	1/2 - 2 t
247	G	Norsk	4	Ganske bra	Oftest pc	Hver dag	1/2 - 2 t
248	J	Norsk	4	Bra	Alltid pc	Hver dag	1/2 - 2 t
249	J	Norsk	4	Ganske bra	Oftest pc	Hver dag	over 2 t
251	G	Norsk	4	Ganske bra	Alltid pc	Hver dag	over 2 t
252	G	Norsk	6	Bra	Alltid pc	Hver dag	over 2 t
259	G	Norsk	4	Ganske bra	Alltid pc	Hver dag	over 2 t
260	J	Norsk	5	Bra	Alltid pc	Hver dag	1/2 - 2 t
261	J	Norsk	6	Bra	Alltid pc	Hver dag	over 2 t
264	J	Norsk	5	Ganske bra	Oftest pc	Hver dag	1/2 - 2 t
265	G	Norsk	3	Ganske bra	Oftest pc	Hver dag	over 2 t
266	G	Norsk	6	Bra	Oftest pc	Hver dag	over 2 t
269	J	Norsk	5	Ganske bra	Oftest pc	Hver dag	over 2 t
272	G	Norsk	5	Ganske bra	Oftest hånd	Hver dag	1/2 - 2 t
273	G	Norsk	6	Bra	Oftest pc	Hver dag	over 2 t
278	G	Flere	4	Bra	Alltid pc	Hver dag	over 2 t
281	J	Norsk	5	Ganske bra	Oftest pc	Hver dag	over 2 t
285	J	Norsk	4	Bra	Oftest hånd	Hver dag	1/2 - 2 t
292	J	Norsk	4	Ganske bra	Oftest hånd	Hver dag	1/2 - 2 t
293	J	Norsk	4	Bra	Oftest pc	Hver dag	over 2 t
294	J	Norsk	4	Bra	Oftest pc	Hver dag	over 2 t
296	G	Norsk	4	Bra	Alltid pc	Hver dag	1/2 - 2 t
297	G	Norsk	5	Bra	Alltid pc	1-4 dager	over 2 t
298	J	Norsk	5	Bra	Alltid pc	Hver dag	1/2 - 2 t
301	J	Norsk	4	Ganske bra	Alltid pc	Hver dag	over 2 t
302	J	Norsk	5	Ganske bra	Oftest pc	Hver dag	over 2 t
303	G	Norsk	4	Bra	Oftest pc	Hver dag	1/2 - 2 t
305	J	Norsk	5	Bra	Oftest pc	1-4 dager	0-30 min
307	J	Norsk	4	Bra	Oftest hånd	Hver dag	over 2 t
309	G	Norsk	5	Bra	Oftest pc	1-4 dager	1/2 - 2 t
310	J	Norsk	5	Bra	Oftest pc	1-4 dager	over 2 t
312	G	Norsk	5	Bra	Alltid pc	Hver dag	over 2 t
313	J	Norsk	4	Bra	Oftest pc	Sjeldnere	over 2 t
317	J	Norsk	5	Ganske bra	Oftest hånd	Hver dag	over 2 t

Spørsmål som handler om hvorfor eleven foretrekker å bruke pc til skolearbeid.

ID	Raskere	Penere	Tenke	Endre	Tegne	Stave	Kopiere	Samarbeide
202	Litt viktig	Viktig	Ikke viktig	Viktig	Usant	Litt viktig	Usant	Ikke viktig
203	Litt viktig	Litt viktig	Usant	Viktig	Usant	Viktig	Usant	Usant
205	Litt viktig	Ikke viktig	Ikke viktig	Viktig	Usant	Litt viktig	Usant	Ikke viktig
206	Ikke viktig	Viktig	Litt viktig	Viktig	Viktig	Viktig	Litt viktig	Ikke viktig
208	Viktig	Litt viktig	Litt viktig	Viktig	Ikke viktig	Litt viktig	Usant	Usant
210	Litt viktig	Viktig	Viktig	Litt viktig	Ikke viktig	Ikke viktig	Usant	Usant
211	Viktig	Ikke viktig	Litt viktig	Viktig	Viktig	Viktig	Viktig	Viktig
212	Viktig	Viktig	Litt viktig	Viktig	Usant	Viktig	Ikke viktig	Viktig
213	Viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Viktig	Ikke viktig	Litt viktig
215	Viktig	Viktig	Usant	Viktig	Usant	Viktig	Litt viktig	Viktig
218	Viktig	Ikke viktig	Litt viktig	Viktig	Ikke viktig	Litt viktig	Ikke viktig	Litt viktig
222	Viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Viktig	Litt viktig	Viktig
225	Litt viktig	Viktig	Usant	Viktig	Litt viktig	Ikke viktig	Ikke viktig	Viktig
226	Litt viktig	Viktig	Litt viktig	Viktig	Litt viktig	Viktig	Litt viktig	Litt viktig
227	Ikke viktig	Ikke viktig	Ikke viktig	Viktig	Usant	Usant	Usant	Ikke viktig
232	Viktig	Viktig	Usant	Viktig	Usant	Litt viktig	Usant	Usant
233	Viktig	Viktig	Usant	Viktig	Usant	Viktig	Ikke viktig	Viktig
234	Litt viktig	Ikke viktig	Viktig	Viktig	Ikke viktig	Ikke viktig	Ikke viktig	Ikke viktig
235	Litt viktig	Ikke viktig	Ikke viktig	Viktig	Ikke viktig	Viktig	Litt viktig	Litt viktig
236	Litt viktig	Litt viktig	Ikke viktig	Viktig	Ikke viktig	Viktig	Ikke viktig	Ikke viktig
237	Litt viktig	Viktig	Viktig	Viktig	Ikke viktig	Viktig	Usant	Litt viktig
238	Viktig	Viktig	Viktig	Viktig	Litt viktig	Viktig	Viktig	Viktig
239	Viktig	Viktig	Litt viktig	Viktig	Viktig	Litt viktig	Viktig	Viktig
241	Viktig	Viktig	Ikke viktig	Viktig	Viktig	Viktig	Viktig	Usant
242	Viktig	Viktig	Viktig	Viktig	Ikke viktig	Viktig	Ikke viktig	Viktig
243	Viktig	Litt viktig	Usant	Viktig	Usant	Litt viktig	Usant	Litt viktig
245	Viktig	Litt viktig	Ikke viktig	Viktig	Ikke viktig	Litt viktig	Litt viktig	Viktig
247	Litt viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Litt viktig	Ikke viktig	Litt viktig
248	Viktig	Litt viktig	Viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Litt viktig
249	Litt viktig	Litt viktig	Usant	Viktig	Ikke viktig	Ikke viktig	Usant	Litt viktig
251	Viktig	Viktig	Viktig	Viktig	Usant	Ikke viktig	Usant	Litt viktig
252	Litt viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Ikke viktig	Usant	Ikke viktig
259	Viktig	Viktig	Litt viktig	Viktig	Viktig	Viktig	Litt viktig	Litt viktig
260	Viktig	Viktig	Viktig	Viktig	Viktig	Viktig	Litt viktig	Usant
261	Viktig	Viktig	Usant	Viktig	Usant	Litt viktig	Litt viktig	Litt viktig
264	Litt viktig	Viktig	Litt viktig	Viktig	Usant	Litt viktig	Ikke viktig	Litt viktig
265	Viktig	Viktig	Viktig	Viktig	Viktig	Viktig	Ikke viktig	Litt viktig
266	Viktig	Litt viktig	Viktig	Viktig	Ikke viktig	Litt viktig	Usant	Ikke viktig
269	Viktig	Viktig	Viktig	Viktig	Litt viktig	Litt viktig	Ikke viktig	Viktig
272	NA	NA	NA	NA	NA	NA	NA	NA
273	Viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Litt viktig	Litt viktig	Litt viktig
278	Litt viktig	Litt viktig	Ikke viktig	Litt viktig	Litt viktig	Ikke viktig	Ikke viktig	Ikke viktig
281	Viktig	Viktig	Viktig	Viktig	Litt viktig	Viktig	Usant	Litt viktig
285	NA	NA	NA	NA	NA	NA	NA	NA
292	NA	NA	NA	NA	NA	NA	NA	NA
293	Viktig	Viktig	Litt viktig	Viktig	Usant	Ikke viktig	Litt viktig	Litt viktig
294	Viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Litt viktig	Ikke viktig	Ikke viktig
296	Litt viktig	Viktig	Litt viktig	Viktig	Viktig	Ikke viktig	Litt viktig	Viktig
297	Litt viktig	Litt viktig	Ikke viktig	Litt viktig	Litt viktig	Litt viktig	Ikke viktig	Ikke viktig
298	Litt viktig	Ikke viktig	Ikke viktig	Viktig	Usant	Viktig	Ikke viktig	Ikke viktig
301	Viktig	Litt viktig	Viktig	Viktig	Ikke viktig	Ikke viktig	Litt viktig	Viktig
302	Viktig	Viktig	Viktig	Viktig	Ikke viktig	Ikke viktig	Litt viktig	Viktig
303	Viktig	Ikke viktig	Ikke viktig	Viktig	Litt viktig	Usant	Usant	Litt viktig
305	Litt viktig	Litt viktig	Usant	Viktig	Usant	Litt viktig	Ikke viktig	Ikke viktig

307	Litt viktig	Litt viktig	Ikke viktig	Litt viktig	Usant	Ikke viktig	Ikke viktig	Ikke viktig
309	Viktig	Viktig	Litt viktig	Viktig	Litt viktig	Viktig	Ikke viktig	Litt viktig
310	Viktig	Viktig	Ikke viktig	Viktig	Litt viktig	Viktig	Viktig	Viktig
312	Viktig	Viktig	Viktig	Viktig	Viktig	Litt viktig	Ikke viktig	Ikke viktig
313	Litt viktig	Ikke viktig	Viktig	Viktig	Ikke viktig	Ikke viktig	Usant	Ikke viktig
317	NA	NA	NA	NA	NA	NA	NA	NA

Spørsmål om hva eleven bruker pc til hjemme.

ID	Skriftlig skolearbeid	Sosial skriving	Annen skriving	Muntlig	Surfing	Spill	Annet
202	En del	Mye	Lite	Aldri	En del	Lite	
203	Mye	En del	Lite	Aldri	Mye	Lite	
205	En del	Mye	En del	Lite	En del	Lite	
206	En del	Mye	Lite	Aldri	Mye	Mye	
208	En del	Mye	En del	Lite	Mye	En del	
210	En del	Mye	En del	Lite	Mye	Mye	Mye
211	En del	Mye	Mye	Lite	Mye	Mye	Mye
212	Mye	Mye	Lite	Lite	En del	Lite	
213	Mye	Mye	Lite	Lite	En del	Lite	
215	Lite	Mye	Lite	Mye	Mye	Mye	En del
218	Mye	Mye	En del	Aldri	En del	En del	Lite
222	Mye	Mye	En del	Lite	En del	En del	Aldri
225	En del	Mye	En del	Aldri	Mye	Lite	En del
226	En del	Mye	En del	Lite	Mye	Lite	Lite
227	En del	Lite	Mye	Aldri	Lite	Lite	
232	En del	Mye	Lite	En del	Mye	Mye	Aldri
233	En del	Mye	Lite	Lite	En del	En del	Lite
234	En del	Mye	Lite	Mye	Lite	Aldri	
235	En del	Mye	En del	Aldri	Mye	En del	
236	En del	En del	En del	Lite	En del	Lite	En del
237	Mye	Mye		Mye	Mye	Lite	
238	Mye	Mye	Mye	Mye	Mye	Mye	
239	Lite	Mye	Lite	En del	Mye	Mye	
241	Lite	Mye	En del	Mye	En del	Mye	En del
242	Lite	Mye	Lite	Mye	Mye	Mye	Mye
243	En del	Mye	Aldri	Aldri	Lite	Lite	
245	Mye	En del	Lite	Lite	Mye	En del	Lite
247	Lite	Mye	En del	Mye	Mye	En del	En del
248	En del	Mye	Lite	Lite	Mye	En del	Aldri
249	En del	Mye	Lite	Aldri	Mye	En del	Aldri
251	En del	En del	Lite	Lite	Mye	En del	Lite
252	En del	Mye	Mye	Mye	Mye	Mye	
259	Mye	Mye	Mye	Mye	Mye	En del	Mye
260	Mye	En del	Mye	Aldri	Mye	Lite	
261	En del	Mye	Mye	Lite	Mye	En del	Mye
264	En del	En del	En del	Aldri	En del	Lite	En del
265	En del	Mye	En del	Lite	Mye	En del	Lite
266	En del	En del	Lite	Lite	Mye	Mye	
269	En del	Mye	Mye	Lite	Mye	En del	Mye

272	Lite	En del	Aldri	Aldri	En del	Lite	Aldri
273	En del	Mye	En del	Lite	Mye	Lite	Mye
278	En del	Mye	En del	En del	Mye	Mye	
281	En del	Mye	En del	En del	Mye	Lite	En del
285	Mye	Mye	En del	Aldri	Mye	Lite	
292	Aldri	Lite	Lite	Aldri	En del	Aldri	En del
293	En del	Mye	Lite	Aldri	Mye	En del	Mye
294	Lite	Mye	Lite	Aldri	Mye	En del	
296	Lite	Mye	Lite	Lite	En del	En del	En del
297	Lite	Mye	Lite	Mye	Mye	Mye	En del
298	Mye	En del	Lite	Lite	En del	Lite	
301	En del	Mye	En del	En del	En del	En del	Aldri
302	En del	Mye	En del	En del	En del	Mye	
303	En del	Mye	Lite	Mye	Mye	En del	
305	En del	Lite	Lite	Aldri	En del	Aldri	Lite
307	Lite	En del	En del	Aldri	En del	Lite	En del
309	Lite	Mye	En del	Lite	Mye	Lite	Mye
310	En del	Mye	En del	En del	Mye	En del	
312	En del	Mye	En del	Aldri	Lite	Mye	Lite
313	Lite	Mye	En del	Lite	Mye	Aldri	
317	Lite	Mye	Lite	En del	Mye	En del	

D3. Avrundede tekstverdier

Verdier for de 13 sentrale leksikosyntaktiske variablene avrundet til et passe antall desimaler for oversikt over de enkelte verdiene.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
					MOSTTR				korte subklaususer					subklaususfrek
		ordlengde		leks.tetthet		T-enhetslengde			preposisjoner		korte-forfelt			
Tekst		Verktøy	leksordlengde		log-TTR		Klaususlengde				adv.subkl		attr.adj	
A2-202	H	3,87	5,84	0,34	0,82	1,31	13,1	6,45	0,011	0,62	0,055	0,69	0,008	1,03
A1-202	T	4,01	5,55	0,4	0,64	1,26	11,8	6,3	0,005	0,72	0,045	0,64	0,018	0,87
A1-203	H	4,46	6,09	0,48	0,76	1,37	9,9	6,26	0,009	0,59	0,044	0,58	0,033	0,58
A2-203	T	4,22	6,05	0,4	0,77	1,35	13,4	7,49	0,009	1	0,034	0,52	0,02	0,79
A2-205	H	4,09	6,19	0,37	0,72	1,28	17	6,49	0,007	0,84	0,127	0,61	0,012	1,62
A1-205	T	4,2	5,78	0,44	0,52	1,24	10	6,25	0,004	0,72	0,042	0,77	0,013	0,60
A1-206	H	4,15	5,77	0,41	0,69	1,33	10,4	7,26	0	0,94	0,1	0,71	0,026	0,43
A2-206	T	4,33	6,19	0,41	0,78	1,36	13,7	6,93	0,008	1,03	0,147	0,54	0,019	0,97
A2-208	H	3,93	5,87	0,37	0,74	1,3	15,7	6,59	0,011	0,64	0,24	0,55	0,033	1,38
A1-208	T	4,13	5,8	0,4	0,72	1,26	14,9	7,93	0,004	0,84	0,133	0,58	0,027	0,88
A1-210	H	4,17	5,75	0,42	0,8	1,4	14,5	7,6	0,008	1,04	0,107	0,56	0,033	0,91
A2-210	T	4,1	5,99	0,38	0,88	1,43	19	8,15	0,001	1,23	0,137	0,58	0,038	1,33
A2-211	H	4,29	6,07	0,42	0,79	1,38	17,2	7,41	0,006	1,11	0,068	0,65	0,028	1,32
A1-211	T	4,53	6,45	0,44	0,75	1,35	18	6,35	0,014	0,76	0,088	0,33	0,032	1,83
A1-212	H	4,2	5,66	0,42	0,67	1,32	17,1	7,81	0,002	1,11	0,088	0,35	0,013	1,19
A2-212	T	4,19	6,14	0,39	0,72	1,33	14,5	6,6	0,01	0,88	0,104	0,55	0,03	1,20
A2-213	H	4,34	6,25	0,44	0,7	1,31	10,7	6,15	0,007	0,64	0,045	0,73	0,022	0,74
A1-213	T	4,4	6,21	0,45	0,76	1,35	12,2	6,94	0,008	0,9	0,087	0,62	0,026	0,76
A2-215	H	4,28	6,26	0,41	0,83	1,38	14	8,37	0,007	1,06	0,057	0,58	0,024	0,67
A1-215	T	4,38	6,53	0,37	0,74	1,32	12	6,66	0,012	0,87	0,079	0,52	0,02	0,81
A1-218	H	4,37	6,5	0,38	0,78	1,3	16,2	7,08	0,002	0,94	0,078	0,46	0,024	1,29
A2-218	T	4,26	6,32	0,36	0,79	1,34	14,7	6,84	0,006	0,68	0,095	0,44	0,023	1,15
A1-222	H	4,2	5,98	0,44	0,75	1,37	13,5	7,83	0	1,12	0,092	0,54	0,024	0,73
A2-222	T	4,08	5,98	0,38	0,76	1,34	17,5	7,31	0,003	0,96	0,175	0,47	0,025	1,40
A1-225	H	4	5,71	0,37	0,67	1,24	11,3	7,21	0,008	0,85	0	0,73	0,017	0,57
A2-225	T	4,04	5,83	0,36	0,65	1,29	11,9	6,34	0,015	0,7	0,137	0,5	0,011	0,87
A2-226	H	4,14	5,78	0,41	0,69	1,3	19	6,82	0,011	1	0,128	0,18	0,011	1,79
A1-226	T	4,16	6,09	0,37	0,71	1,29	22,4	8,35	0,011	1,07	0,14	0,46	0,011	1,69

A1-227	H	4,52	6,5	0,42	0,84	1,38	14,9	6,45	0,01	0,78	0,183	0,79	0,023	1,31
A2-227	T	4,51	6,71	0,4	0,82	1,4	18,2	6,87	0,008	0,91	0,174	0,46	0,034	1,65
A2-232	H	4	5,59	0,35	0,72	1,33	13,7	5,53	0,01	0,7	0,175	0,61	0,016	1,48
A1-232	T	4,04	5,55	0,37	0,66	1,3	14,3	5,67	0,014	0,68	0,127	0,56	0,006	1,52
A1-233	H	4,18	6,15	0,4	0,73	1,38	11,4	7,38	0,003	1,23	0,111	0,76	0,015	0,55
A2-233	T	3,7	5,32	0,35	0,77	1,35	12,2	6,34	0,005	0,78	0,131	0,69	0,011	0,93
A1-234	H	4,1	5,71	0,42	0,72	1,33	11,7	7,21	0,004	1,03	0,059	0,57	0,02	0,62
A2-234	T	3,86	5,7	0,34	0,8	1,33	17,1	6,35	0,001	0,73	0,146	0,61	0,017	1,69
A2-235	H	4,2	6,14	0,38	0,8	1,4	11,9	6,75	0,006	0,8	0,118	0,62	0,02	0,76
A1-235	T	4,24	6,4	0,37	0,71	1,28	16,3	6,22	0,015	0,58	0,127	0,67	0,02	1,62
A2-236	H	4,4	6,49	0,41	0,84	1,38	16	8,15	0,005	1,28	0,094	0,38	0,032	0,96
A1-236	T	4,41	6,57	0,42	0,76	1,36	18,3	7,56	0,002	1,06	0,048	0,42	0,021	1,42
A1-237	H	4,24	6,12	0,42	0,72	1,32	12,5	6,85	0,012	0,98	0,065	0,45	0,019	0,82
A2-237	T	4,61	6,53	0,46	0,73	1,38	12,7	7,67	0,005	0,99	0,068	0,45	0,023	0,66
A1-238	H	3,77	5,62	0,36	0,8	1,28	12,2	8,24	0,004	1,03	0,029	0,43	0,018	0,48
A2-238	T	3,99	5,96	0,37	0,78	1,33	11,7	6,63	0,002	0,72	0,122	0,55	0,008	0,76
A1-239	H	4,21	5,78	0,45	0,68	1,3	11	7,21	0,011	0,92	0,105	0,75	0,022	0,52
A2-239	T	4,1	5,82	0,39	0,8	1,38	13,1	6,67	0,009	0,85	0,112	0,68	0,021	0,96
A1-241	H	4,03	5,81	0,39	0,76	1,35	15,3	7,28	0,014	0,92	0,125	0,53	0,01	1,11
A2-241	T	4,2	6,2	0,39	0,73	1,31	16,2	7,72	0,002	1,11	0,169	0,55	0,032	1,10
A2-242	H	4,1	6,05	0,38	0,83	1,35	13,9	5,65	0,007	0,69	0,13	0,5	0,02	1,45
A1-242	T	4,13	6,04	0,37	0,8	1,33	14,4	6,93	0,008	0,94	0,148	0,56	0,008	1,08
A2-243	H	4,13	5,91	0,39	0,67	1,27	14,8	6,45	0,009	1,01	0,183	0,5	0,015	1,29
A1-243	T	3,85	5,33	0,38	0,44	1,16	13,1	6,48	0,015	0,79	0,028	0,63	0	1,03
A2-245	H	4,18	6,03	0,39	0,78	1,31	16,1	6,51	0,008	0,9	0,19	0,52	0,018	1,47
A1-245	T	4,09	5,66	0,39	0,66	1,3	16,6	6,88	0,01	0,89	0,187	0,61	0,004	1,42
A1-247	H	3,97	5,58	0,4	0,53	1,21	16	6,47	0,005	0,93	0,14	0,7	0,014	1,48
A2-247	T	3,98	5,83	0,36	0,72	1,29	19,2	7,25	0,007	1,03	0,262	0,52	0,016	1,65
A2-248	H	3,93	5,88	0,36	0,63	1,27	15	5,88	0,006	0,64	0,179	0,59	0,003	1,55
A1-248	T	3,86	5,17	0,38	0,6	1,25	14,8	6,29	0,008	0,46	0,034	0,84	0,003	1,36
A1-249	H	3,89	5,27	0,39	0,55	1,2	12,5	6,23	0,014	0,73	0,071	0,65	0,006	1,00
A2-249	T	3,71	5,36	0,31	0,67	1,24	14,3	5,86	0,008	0,67	0,171	0,59	0,005	1,43
A2-251	H	4,11	6,03	0,37	0,81	1,35	13,3	6,81	0	0,78	0,216	0,74	0,008	0,95
A1-251	T	4,19	5,85	0,41	0,76	1,38	13,5	7,79	0,01	0,92	0,077	0,66	0,025	0,73
A2-252	H	4,23	6,41	0,38	0,8	1,35	17,9	7,64	0,008	1,06	0,149	0,28	0,028	1,35
A1-252	T	4,37	6,37	0,36	0,77	1,34	17,2	6,44	0,012	0,73	0,219	0,36	0,019	1,67
A2-259	H	4,46	6,49	0,41	0,67	1,31	12,7	6,25	0,008	0,7	0,06	0,5	0,019	1,02
A1-259	T	4,01	5,76	0,35	0,62	1,27	15,3	7,3	0,009	0,81	0,143	0,48	0,004	1,10
A1-260	H	4,71	6,83	0,44	0,88	1,4	13	7,47	0,008	1,01	0,059	0,59	0,024	0,74
A2-260	T	4,67	6,9	0,42	0,86	1,43	17,1	7,57	0,005	1,05	0,117	0,46	0,029	1,27
A1-261	H	4,06	5,65	0,4	0,8	1,35	20,3	7,57	0,005	0,9	0,078	0,53	0,026	1,68
A2-261	T	4,33	6,3	0,4	0,83	1,38	15,2	7,17	0,005	0,81	0,134	0,48	0,025	1,11
A1-264	H	4,18	5,85	0,39	0,76	1,36	9,6	6,29	0,024	0,78	0,103	0,65	0,011	0,53
A2-264	T	3,99	5,73	0,38	0,77	1,33	13	7,16	0,006	0,98	0,169	0,67	0,013	0,82
A1-265	H	4,35	6,48	0,42	0,75	1,3	10,1	8,03	0,004	1,09	0	0,73	0,029	0,26
A2-265	T	4,31	6,27	0,4	0,82	1,36	17,2	7,97	0,006	1,19	0,075	0,38	0,036	1,16
A1-266	H	4,29	5,97	0,46	0,81	1,35	13,4	7,33	0,011	1,08	0,031	0,52	0,038	0,83
A2-266	T	4,3	6,2	0,4	0,85	1,42	14,2	7,36	0,008	0,98	0,136	0,44	0,03	0,93
A1-269	H	4,51	6,4	0,43	0,75	1,38	21	8,29	0	1,26	0,079	0,45	0,019	1,53
A2-269	T	4,14	5,9	0,36	0,82	1,36	19,2	6,6	0,003	0,79	0,146	0,48	0,013	1,91
A1-272	H	4,52	6,3	0,42	0,64	1,3	11,8	7,1	0,017	0,78	0,1	0,69	0,014	0,67
A2-272	T	4,16	6,07	0,4	0,78	1,36	16	7,25	0,002	0,87	0,214	0,53	0,028	1,21
A2-273	H	4,54	6,59	0,43	0,74	1,31	15,8	7,52	0,002	1,13	0,104	0,5	0,014	1,09
A1-273	T	4,73	6,58	0,45	0,76	1,39	16,2	8,38	0,004	1,3	0,141	0,47	0,035	0,94
A2-278	H	4,42	6,29	0,43	0,74	1,33	14,4	7,33	0,006	0,88	0,163	0,62	0,028	0,96
A1-278	T	4,29	6,07	0,38	0,67	1,31	13,1	6,21	0,011	0,63	0,035	0,58	0,011	1,11
A2-281	H	4,27	6,2	0,41	0,73	1,33	13,2	6,43	0,008	0,85	0,12	0,44	0,012	1,04
A1-281	T	4,15	5,73	0,44	0,67	1,28	19,5	8,14	0,01	1,14	0,107	0,49	0,019	1,40
A2-285	H	4,08	6,02	0,37	0,86	1,39	11,4	6,28	0,006	0,77	0,118	0,71	0,017	0,82
A1-285	T	3,96	5,55	0,37	0,72	1,33	12	6,79	0,012	0,79	0,072	0,64	0,018	0,76
A1-292	H	4,15	5,86	0,38	0,69	1,27	12	6,19	0,014	0,69	0,074	0,56	0,012	0,94
A2-292	T	4,25	6,33	0,38	0,8	1,32	13,7	6,78	0,004	0,74	0,169	0,62	0,029	1,03
A1-293	H	3,99	5,56	0,39	0,76	1,34	13,6	6,43	0,014	0,86	0,149	0,65	0,007	1,12
A2-293	T	3,91	5,64	0,34	0,73	1,3	16,4	7,21	0,006	1,03	0,173	0,61	0,013	1,27
A2-294	H	4,01	5,96	0,36	0,81	1,34	14,3	6,62	0,008	0,96	0,054	0,55	0,013	1,15
A1-294	T	3,86	5,43	0,39	0,62	1,26	13,3	6,36	0,009	0,73	0,068	0,37	0,015	1,09
A2-296	H	4,01	6,02	0,34	0,76	1,32	15,5	7,1	0,01	0,7	0,214	0,53	0,014	1,19
A1-296	T	4,01	5,83	0,38	0,71	1,28	17,7	7,28	0,014	0,88	0,118	0,39	0,02	1,43
A2-297	H	4,24	5,89	0,41	0,85	1,4	14,2	6,91	0	0,9	0,157	0,61	0,014	1,06
A1-297	T	4,52	6,3	0,44	0,85	1,43	16,3	8,16	0,004	1,16	0,21	0,48	0,032	1,00
A2-298	H	4,25	6,24	0,39	0,81	1,4	13,9	6,67	0,012	0,92	0,135	0,52	0,017	1,08
A1-298	T	4,3	5,67	0,46	0,66	1,32	13,1	6,57	0,005	0,86	0	0,79	0,033	1,00
A1-301	H	4,45	6,47	0,39	0,7	1,35	12,8	7,48	0,002	0,97	0,06	0,44	0,016	0,72
A2-301	T	4,51	6,42	0,44	0,8	1,35	12,4	7,68	0,001	0,97	0,138	0,57	0,028	0,61
A2-302	H	4,23	6,3	0,4	0,72	1,36	15,4	6,43	0	0,97	0,15	0,48	0,013	1,40
A1-302	T	4,15	5,91	0,41	0,66	1,32	14,9	7,03	0,006	0,79	0,059	0,65	0,01	1,12

A1-303	H	4,34	6,12	0,39	0,72	1,32	14,8	7,76	0	1,05	0,105	0,47	0,02	0,90
A2-303	T	4,48	6,44	0,45	0,82	1,38	15	7,48	0,002	1,04	0,074	0,44	0,037	1,00
A2-305	H	4,05	5,82	0,39	0,72	1,34	14,4	6,51	0,011	0,91	0,159	0,69	0,009	1,22
A1-305	T	3,99	5,6	0,36	0,56	1,23	12,4	5,63	0,017	0,73	0,036	0,62	0,004	1,21
A1-307	H	4,21	5,45	0,48	0,64	1,27	10,8	6,96	0	0,89	0,143	0,56	0,01	0,56
A2-307	T	4,03	5,94	0,38	0,64	1,26	13,7	6,62	0,005	0,57	0,167	0,67	0,008	1,07
A2-309	H	3,95	5,78	0,39	0,8	1,31	13,2	6,33	0,002	0,78	0,11	0,62	0,019	1,09
A1-309	T	4,28	5,91	0,43	0,74	1,34	12,3	6,97	0,005	0,84	0	0,59	0,025	0,77
A2-310	H	4,39	6,53	0,41	0,89	1,4	13,9	6,58	0,008	0,92	0,056	0,75	0,017	1,12
A1-310	T	4,38	5,99	0,41	0,81	1,36	15,2	7,33	0,013	1,06	0,096	0,64	0,021	1,08
A2-312	H	4,27	6,41	0,38	0,88	1,4	19,8	7,28	0,01	0,82	0,158	0,44	0,031	1,71
A1-312	T	4,23	5,97	0,39	0,76	1,37	15,8	7,62	0,01	0,95	0,087	0,48	0,026	1,07
A1-313	H	4,28	5,83	0,44	0,71	1,32	14,7	7,2	0,003	0,87	0,067	0,5	0,015	1,05
A2-313	T	4,09	6,06	0,36	0,79	1,34	13	6,48	0,004	0,79	0,096	0,64	0,013	1,00
A1-317	H	4,56	6,32	0,42	0,7	1,31	13,4	7,46	0,015	0,95	0,033	0,61	0,007	0,79
A2-317	T	4,72	6,72	0,45	0,77	1,39	15	7,57	0,004	1,08	0,081	0,58	0,028	0,98

D4. Nøyaktige tekstverdier

Verdier for de 13 sentrale leksikosyntaktiske variablene avrundet til 8 desimaler. Tabellen kan kopieres fra en elektronisk versjon av avhandlingen inn i et statistikkprogram for å verifisere utregningene.

id	M	ordlengde	leksordL	leks_tetthet	MOSTTR	logTTR	T-enhetsL	klaususL	korte_subkl	prep_kl	advkl_kl	klaus_Te	korte_forf	attr_adj
A2-202	H	3,86624204	5,84375	0,33970276	0,825	1,30960776	13,08333333	6,45205479	0,01061571	0,61643836	0,05479452	1,02777778	0,68571429	
		0,00849257												
A1-202	T	4,00722022	5,54751131	0,39891697	0,64	1,25501769	11,78723404	6,29545455	0,00541516	0,71590909	0,04545455	0,87234043	0,63829787	
		0,01805054												
A1-203	H	4,45539906	6,08866995	0,47652582	0,76	1,36560398	9,90697674	6,26470588	0,00938967	0,58823529	0,04411765	0,58139535	0,57894737	
		0,03286385												
A2-203	T	4,21719457	6,04545455	0,39819005	0,77	1,34708045	13,39393939	7,49152542	0,00904977	1	0,03389831	0,78787879	0,51612903	
		0,02036199												
A2-205	H	4,08801956	6,18543046	0,36919315	0,72	1,27662759	17,04166667	6,49206349	0,00733496	0,84126984	0,12698413	1,625	0,60869565	
		0,01222494												
A1-205	T	4,19555556	5,77777778	0,44	0,525	1,23949527	10	6,25	0,00444444	0,72222222	0,04166667	0,6	0,76744186	
		0,01333333												
A1-206	H	4,1515748	5,76699029	0,40551181	0,692	1,33269947	10,36734694	7,25714286	0	0,94285714	0,1	0,42857143	0,71428571	
		0,02559055												
A2-206	T	4,33461538	6,18867925	0,40769231	0,776	1,35732115	13,68421053	6,93333333	0,00769231	1,02666667	0,14666667	0,97368421	0,54054054	
		0,01923077												
A2-208	H	3,92716237	5,86885246	0,37025797	0,736	1,29860574	15,69047619	6,59	0,01062215	0,64	0,24	1,38095238	0,54761905	
		0,03338392												
A1-208	T	4,13165266	5,79858657	0,39635854	0,72333333	1,26018125	14,875	7,93333333	0,00420168	0,84444444	0,13333333	0,875	0,58333333	
		0,02661064												
A1-210	H	4,169279	5,74716981	0,4153605	0,8	1,39603471	14,5	7,5952381	0,00783699	1,03571429	0,10714286	0,90909091	0,55555556	
		0,03291536												
A2-210	T	4,10114633	5,98926655	0,37693864	0,88333333	1,43381635	19,01282051	8,14835165	0,00134862	1,22527473	0,13736264	1,33333333	0,57746479	0,0384356
A2-211	H	4,28834356	6,07246377	0,42331288	0,79333333	1,3765739	17,15789474	7,40909091	0,00613497	1,11363636	0,06818182	1,31578947	0,64705882	
		0,02760736												
A1-211	T	4,52777778	6,45263158	0,43981481	0,75	1,35161448	18	6,35294118	0,01388889	0,76470588	0,08823529	1,83333333	0,33333333	
		0,03240741												
A1-212	H	4,19550562	5,6631016	0,42022472	0,67	1,31807488	17,11538462	7,80701754	0,00224719	1,10526316	0,0877193	1,19230769	0,34782609	
		0,01348315												
A2-212	T	4,19488189	6,14	0,39370079	0,725	1,33466797	14,51428571	6,5974026	0,00984252	0,88311688	0,1038961	1,2	0,54545455	
		0,02952756												
A2-213	H	4,34236453	6,24719101	0,43842365	0,7	1,31291703	10,68421053	6,15151515	0,00738916	0,63636364	0,04545455	0,73684211	0,72972973	
		0,02216749												
A1-213	T	4,40304709	6,20552147	0,45152355	0,76285714	1,35137615	12,23728814	6,94230769	0,00831025	0,90384615	0,08653846	0,76271186	0,61538462	
		0,02631579												
A2-215	H	4,28327645	6,2605042	0,40614334	0,83333333	1,37898251	13,95238095	8,37142857	0,00682594	1,05714286	0,05714286	0,66666667	0,57894737	
		0,02389078												
A1-215	T	4,38339921	6,52688172	0,36758893	0,74	1,31668507	12,04761905	6,65789474	0,01185771	0,86842105	0,07894737	0,80952381	0,52380952	
		0,01976285												
A1-218	H	4,37086093	6,5	0,37969095	0,775	1,30223958	16,17857143	7,078125	0,00220751	0,9375	0,078125	1,28571429	0,46428571	
		0,02428256												
A2-218	T	4,25725095	6,32042254	0,35813367	0,79333333	1,34483182	14,68518519	6,8362069	0,00630517	0,68103448	0,09482759	1,14814815	0,44230769	
		0,02269861												
A1-222	H	4,19831933	5,98098859	0,44201681	0,74666667	1,36919147	13,52727273	7,82894737	0	1,11842105	0,09210526	0,72727273	0,53658537	
		0,02352941												
A2-222	T	4,07702523	5,97923875	0,38379814	0,75666667	1,33784426	17,51162791	7,31067961	0,00265604	0,96116505	0,17475728	1,39534884	0,47368421	0,0252324
A1-225	H	4,00420168	5,71264368	0,36554622	0,67	1,24268474	11,33333333	7,21212121	0,00840336	0,84848485	0	0,57142857	0,73333333	
		0,01680672												
A2-225	T	4,03671706	5,83233533	0,36069114	0,65	1,29029751	11,87179487	6,34246575	0,01511879	0,69863014	0,1369863	0,87179487	0,5	
		0,01079914												
A2-226	H	4,13909774	5,78181818	0,41353383	0,68666667	1,30328927	19	6,82051282	0,0112782	1	0,12820513	1,78571429	0,18181818	0,0112782
A1-226	T	4,1643454	6,09090909	0,36768802	0,70666667	1,28962949	22,4375	8,34883721	0,01114206	1,06976744	0,13953488	1,6875	0,46153846	
		0,01114206												
A1-227	H	4,5245478	6,50310559	0,41602067	0,845	1,37858084	14,88461538	6,45	0,01033592	0,78333333	0,18333333	1,30769231	0,79166667	
		0,02325581												
A2-227	T	4,51184433	6,71092077	0,39509306	0,824	1,39785841	18,18461538	6,87209302	0,00761421	0,90697674	0,1744186	1,64615385	0,45762712	
		0,0338409												
A2-232	H	3,9968254	5,58715596	0,34603175	0,72	1,3263453	13,69565217	5,52631579	0,00952381	0,70175439	0,1754386	1,47826087	0,60869565	
		0,01587302												
A1-232	T	4,03641457	5,54545455	0,3697479	0,66	1,29958229	14,28	5,66666667	0,0140056	0,68253968	0,12698413	1,52	0,56	
		0,00560224												
A1-233	H	4,17921687	6,14661654	0,40060241	0,73	1,37996755	11,44827586	7,37777778	0,00301205	1,23333333	0,11111111	0,55172414	0,75925926	
		0,01506024												
A2-233	T	3,69785884	5,3166287	0,3481364	0,77333333	1,34859115	12,24271845	6,33668342	0,00475813	0,77889447	0,13065327	0,93203883	0,69230769	0,0111023
A1-234	H	4,09795918	5,71153846	0,4244898	0,72	1,33242691	11,66666667	7,20588235	0,00408163	1,02941176	0,05882353	0,61904762	0,57142857	
		0,02040816												
A2-234	T	3,86436782	5,70169492	0,33908046	0,80333333	1,33104208	17,05882353	6,35036496	0,00114943	0,72992701	0,1459854	1,68627451	0,60784314	
		0,01724138												
A2-235	H	4,2005814	6,13636364	0,38372093	0,8	1,40064474	11,86206897	6,74509804	0,00581395	0,80392157	0,11764706	0,75862069	0,61538462	
		0,02034884												
A1-235	T	4,23976608	6,4047619	0,36842105	0,70666667	1,28306516	16,28571429	6,21818182	0,01461988	0,58181818	0,12727273	1,61904762	0,66666667	
		0,02046784												
A2-236	H	4,39814815	6,48587571	0,40972222	0,84	1,38422558	16	8,1509434	0,00462963	1,28301887	0,09433962	0,96296296	0,38461538	
		0,03240741												
A1-236	T	4,40546218	6,57070707	0,41596639	0,765	1,35944617	18,30769231	7,55555556	0,00210084	1,06349206	0,04761905	1,42307692	0,42307692	0,0210084
A1-237	H	4,24	6,11864407	0,41647059	0,725	1,31931484	12,5	6,85483871	0,01176471	0,98387097	0,06451613	0,82352941	0,45454545	
		0,01882353												
A2-237	T	4,61428571	6,53488372	0,46071429	0,72666667	1,37812255	12,72727273	7,67123288	0,00535714	0,98630137	0,06849315	0,65909091	0,45	
		0,02321429												
A1-238	H	3,775	5,62376238	0,36071429	0,8	1,27963843	12,17391304	8,23529412	0,00357143	1,02941176	0,02941176	0,47826087	0,42857143	
		0,01785714												
A2-238	T	3,99329983	5,95515695	0,37353434	0,784	1,33474896	11,70588235	6,63333333	0,00167504	0,72222222	0,12222222	0,76470588	0,55102041	
		0,00837521												
A1-239	H	4,21167883	5,7804878	0,44890511	0,68	1,30090165	10,96	7,21052632	0,01094891	0,92105263	0,10526316	0,52	0,75	
		0,02189781												
A2-239	T	4,10091743	5,81782946	0,39449541	0,80333333	1,38267508	13,08	6,67346939	0,00917431	0,84693878	0,1122449	0,96	0,68	
		0,02140673												
A1-241	H	4,03436426	5,81415929	0,38831615	0,76	1,35421228	15,31578947	7,275	0,0137457	0,925	0,125	1,10526316	0,52631579	
		0,01030928												
A2-241	T	4,20119522	6,20103093	0,38645418	0,73	1,3100181	16,19354839	7,72307692	0,00199203	1,10769231	0,16923077	1,09677419	0,5483871	
		0,03187251												
A2-242	H	4,09836066	6,05128205	0,38360656	0,82666667	1,3537057	13,86363636	5,64814815	0,00655738	0,68518519	0,12962963	1,45454545	0,5	
		0,01967213												
A1-242	T	4,13368984	6,03571429	0,37433155	0,8	1,33482484	14,38461538	6,92592593	0,00802139	0,94444444	0,14814815	1,07692308	0,56	
		0,00802139												
A2-243	H	4,12663755	5,91061453	0,39082969	0,67	1,27108542	14,77419355	6,45070423	0,00873362	1,01408451	0,18309859	1,29032258	0,5	
		0,01528384												

A1-243	T	3,84565217	5,32758621	0,37826087	0,435	1,16142698	13,14285714	6,47887324	0,01521739	0,78873239	0,02816901	1,02857143	0,62857143	0
A2-245	H	4,18093385	6,02525253	0,38521401	0,775	1,30623348	16,0625	6,50632911	0,0077821	0,89873418	0,18987342	1,46875	0,51612903	0,01750973
A1-245	T	4,08527132	5,65517241	0,39341085	0,664	1,29949069	16,64516129	6,88	0,00968992	0,89333333	0,18666667	1,41935484	0,61290323	0,00387597
A1-247	H	3,97289973	5,58108108	0,40108401	0,53333333	1,20902972	16,04347826	6,47368421	0,00542005	0,92982456	0,14035088	1,47826087	0,69565217	0,01355014
A2-247	T	3,98190045	5,83229814	0,36425339	0,72	1,28599448	19,2173913	7,24590164	0,00678733	1,03278689	0,26229508	1,65217391	0,52173913	0,0158371
A2-248	H	3,92705167	5,88135593	0,35866261	0,63333333	1,26979612	14,95454545	5,875	0,00607903	0,64285714	0,17857143	1,54545455	0,59090909	0,00303951
A1-248	T	3,86253369	5,17142857	0,37735849	0,6	1,24573528	14,84	6,28813559	0,00808625	0,45762712	0,03389831	1,36	0,84	0,00269542
A1-249	H	3,89398281	5,27205882	0,38968481	0,54666667	1,20304592	12,46428571	6,23214286	0,01432665	0,73214286	0,07142857	1	0,65384615	0,00573066
A2-249	T	3,71296296	5,36134454	0,31481481	0,668	1,24439841	14,26415094	5,86046512	0,00793651	0,6744186	0,17054264	1,43396226	0,58695652	0,00529101
A2-251	H	4,10714286	6,03225806	0,36904762	0,81	1,3458717	13,26315789	6,81081081	0	0,78378378	0,21621622	0,94736842	0,73684211	0,00793651
A1-251	T	4,18765432	5,8452381	0,41481481	0,76	1,37780967	13,5	7,78846154	0,00987654	0,92307692	0,07692308	0,73333333	0,65517241	0,02469136
A2-252	H	4,23119777	6,41304348	0,38440111	0,8	1,35369634	17,95	7,63829787	0,00835655	1,06382979	0,14893617	1,35	0,27777778	0,02785515
A1-252	T	4,37378641	6,36666667	0,36407767	0,76666667	1,33553469	17,16666667	6,4375	0,01213592	0,734375	0,21875	1,66666667	0,36363636	0,01941748
A2-259	H	4,46242775	6,492891	0,40655106	0,672	1,30778251	12,65853659	6,25301205	0,00770713	0,69879518	0,06024096	1,02439024	0,5	0,01926782
A1-259	T	4,01304348	5,75625	0,34782609	0,615	1,26794872	15,33333333	7,3015873	0,00869565	0,80952381	0,14285714	1,1	0,48275862	0,00434783
A1-260	H	4,71062992	6,83035714	0,44094488	0,876	1,40159711	13,02564103	7,47058824	0,00787402	1,01470588	0,05882353	0,74358974	0,59459459	0,02362205
A2-260	T	4,6702381	6,90340909	0,41904762	0,865	1,42710693	17,14285714	7,56756757	0,0047619	1,04504505	0,11711712	1,26530612	0,45652174	0,02857143
A1-261	H	4,06217617	5,6516129	0,4015544	0,8	1,35359185	20,31578947	7,56862745	0,00518135	0,90196078	0,07843137	1,68421053	0,52631579	0,02590674
A2-261	T	4,32503113	6,29874214	0,39601494	0,83142857	1,38451106	15,1509434	7,16964286	0,00498132	0,8125	0,13392857	1,11320755	0,48	0,0249066
A2-264	H	4,1809872	5,84976526	0,38939671	0,756	1,35602366	9,59649123	6,28735632	0,023766	0,7816092	0,10344828	0,52631579	0,64705882	0,01096892
A2-264	T	3,99058085	5,73360656	0,38304553	0,768	1,32899864	13	7,15730337	0,00627943	0,97752809	0,16853933	0,81632653	0,67391304	0,01255887
A1-265	H	4,35164835	6,47826087	0,42124542	0,74666667	1,29843823	10,11111111	8,02941176	0,003663	1,08823529	0	0,25925926	0,73076923	0,02930403
A2-265	T	4,31460674	6,26976744	0,40262172	0,816	1,35915121	17,22580645	7,97014925	0,00561798	1,19402985	0,07462687	1,16129032	0,37931034	0,03558052
A1-266	H	4,29211087	5,96759259	0,46055437	0,812	1,3512132	13,4	7,328125	0,01066098	1,078125	0,03125	0,82857143	0,51515152	0,03837953
A2-266	T	4,30369128	6,20083682	0,40100671	0,848	1,41737736	14,19047619	7,35802469	0,00838926	0,97530864	0,13580247	0,92857143	0,43589744	0,03020134
A1-269	H	4,50793651	6,40145985	0,43492063	0,74666667	1,38348816	21	8,28947368	0	1,26315789	0,07894737	1,53333333	0,45454545	0,01904762
A2-269	T	4,14037855	5,8961039	0,36435331	0,82	1,35848499	19,21212121	6,60416667	0,00315457	0,79166667	0,14583333	1,90909091	0,48275862	0,0126183
A1-272	H	4,51830986	6,30405405	0,41690141	0,64	1,29556737	11,83333333	7,1	0,01690141	0,78	0,1	0,66666667	0,69230769	0,01408451
A2-272	T	4,16256158	6,06995885	0,39901478	0,78	1,35619534	16,02631579	7,25	0,00164204	0,86904762	0,21428571	1,21052632	0,52631579	0,02791461
A2-273	H	4,53769841	6,58525346	0,43055556	0,744	1,31300662	15,75	7,52238806	0,00198413	1,13432836	0,10447761	1,09375	0,5	0,01388889
A1-273	T	4,72574627	6,57916667	0,44776119	0,764	1,38979694	16,24242424	8,375	0,00373134	1,296875	0,140625	0,93939394	0,46875	0,03544776
A2-278	H	4,4178273	6,29032258	0,43175487	0,735	1,33141222	14,36	7,32653061	0,00557103	0,87755102	0,16326531	0,96	0,625	0,02785515
A1-278	T	4,28813559	6,07407407	0,38135593	0,67333333	1,31331866	13,11111111	6,21052632	0,01129944	0,63157895	0,03508772	1,11111111	0,57692308	0,01129944
A2-281	H	4,26689189	6,2033195	0,40709459	0,728	1,33345689	13,15555556	6,43478261	0,00844595	0,84782609	0,11956522	1,04444444	0,44444444	0,01182432
A1-281	T	4,14766082	5,72666667	0,43859649	0,67	1,28131953	19,54285714	8,14285714	0,01023392	1,14285714	0,10714286	1,4	0,48571429	0,01900585
A2-285	H	4,07644306	6,01666667	0,37441498	0,856	1,38684176	11,44642857	6,28431373	0,00624025	0,7745098	0,11764706	0,82142857	0,70833333	0,01716069
A1-285	T	3,96348645	5,55414013	0,36984688	0,72	1,32982868	11,95774648	6,792	0,01177856	0,792	0,072	0,76056338	0,640625	0,01766784
A1-292	H	4,14726841	5,85534591	0,37767221	0,69	1,27211004	12,02857143	6,19117647	0,01425178	0,69117647	0,07352941	0,94285714	0,55882353	0,01187648
A2-292	T	4,25095785	6,32828283	0,37931034	0,795	1,31789961	13,73684211	6,77922078	0,00383142	0,74025974	0,16883117	1,02631579	0,62162162	0,02873563
A1-293	H	3,98747764	5,56164384	0,39177102	0,764	1,34109093	13,63414634	6,42528736	0,01431127	0,86206897	0,14942529	1,12195122	0,64705882	0,00715564
A2-293	T	3,90757856	5,63783784	0,34195933	0,73	1,29814599	16,39393939	7,21333333	0,00554529	1,02666667	0,17333333	1,27272727	0,61290323	0,012939
A2-294	H	4,01347709	5,96240602	0,35849057	0,81333333	1,33737948	14,26923077	6,625	0,00808625	0,96428571	0,05357143	1,15384615	0,54545455	0,01347709
A1-294	T	3,85560345	5,42777778	0,38793103	0,615	1,25960955	13,25714286	6,35616438	0,00862069	0,7260274	0,06849315	1,08571429	0,36666667	0,01508621
A2-296	H	4,00804829	6,01775148	0,34004024	0,755	1,31854851	15,53125	7,1	0,01006036	0,7	0,21428571	1,1875	0,53125	0,01408451
A1-296	T	4,01414141	5,82795699	0,37575758	0,71	1,27716463	17,67857143	7,27941176	0,01414141	0,88235294	0,11764706	1,42857143	0,39285714	0,02020202
A2-297	H	4,23966942	5,89	0,41322314	0,85	1,40195501	14,23529412	6,91428571	0	0,9	0,15714286	1,05882353	0,60606061	0,01446281
A1-297	T	4,52173913	6,2972973	0,43873518	0,848	1,43455089	16,32258065	8,16129032	0,00395257	1,16129032	0,20967742	1	0,48148148	0,03162055
A2-298	H	4,25072046	6,23880597	0,38616715	0,80666667	1,39724858	13,88	6,67307692	0,01152738	0,92307692	0,13461538	1,08	0,52	0,01729107
A1-298	T	4,29891304	5,66666667	0,45652174	0,66	1,31673428	13,14285714	6,57142857	0,00543478	0,85714286	0	1	0,78571429	0,0326087
A1-301	H	4,45309381	6,4742268	0,38722555	0,695	1,34843593	12,84615385	7,47761194	0,00199601	0,97014925	0,05970149	0,71794872	0,44117647	0,01596806
A2-301	T	4,51347305	6,42424242	0,44461078	0,8	1,35436567	12,37037037	7,67816092	0,00149701	0,96551724	0,13793103	0,61111111	0,56603774	0,02844311
A2-302	H	4,23316062	6,30065359	0,39637306	0,725	1,35877319	15,44	6,43333333	0	0,96666667	0,15	1,4	0,47619048	0,01295337
A1-302	T	4,15271967	5,90721649	0,40585774	0,665	1,31633263	14,9375	7,02941176	0,00627615	0,79411765	0,05882353	1,125	0,64516129	0,01046025
A1-303	H	4,33559322	6,12280702	0,38644068	0,72	1,31867793	14,75	7,76315789	0	1,05263158	0,10526316	0,9	0,46666667	0,02033898
A2-303	T	4,48267327	6,43715847	0,4529703	0,825	1,37619277	14,96296296	7,48148148	0,00247525	1,03703704	0,07407407	1	0,44	0,03712871
A2-305	H	4,04868914	5,81818182	0,39138577	0,716	1,33855196	14,43243243	6,51219512	0,01123596	0,91463415	0,15853659	1,21621622	0,68571429	0,0093633
A1-305	T	3,98731501	5,60465116	0,36363636	0,56	1,23030184	12,44736842	5,63095238	0,01691332	0,72619048	0,03571429	1,21052632	0,62162162	0,00422833
A1-307	H	4,20512821	5,4516129	0,47692308	0,64	1,26867795	10,83333333	6,96428571	0	0,89285714	0,14285714	0,55555556	0,55555556	0,01025641
A2-307	T	4,03274559	5,93959732	0,37531486	0,64	1,25887881	13,68965517	6,61666667	0,00503778	0,56666667	0,16666667	1,06896552	0,66666667	0,00755668
A2-309	H	3,95454545	5,78212291	0,38744589	0,805	1,3059675	13,2	6,32876712	0,0021645	0,78082192	0,10958904	1,08571429	0,625	0,01948052
A1-309	T	4,28240741	5,91304348	0,42592593	0,74	1,33561447	12,34285714	6,96774194	0,00462963	0,83870968	0	0,77142857	0,58823529	0,02546296
A2-310	H	4,39240506	6,53125	0,40506329	0,89	1,39580731	13,94117647	6,58333333	0,00843882	0,91666667	0,05555556	1,11764706	0,75	0,01687764
A1-310	T	4,38057743	5,99363057	0,41207349	0,81	1,36317724	15,24	7,32692308	0,01312336	1,05769231	0,09615385	1,08	0,63636364	0,02099738
A2-312	H	4,2746988	6,41025641	0,37590361	0,885	1,39600752	19,76190476	7,28070175	0,00963855	0,8245614	0,15789474	1,71428571	0,44444444	0,0313253
A1-312	T	4,230837	5,96839729	0,39030837	0,76222222	1,37139115	15,76388889	7,61744966	0,00969163	0,94630872	0,08724832	1,06944444	0,48275862	0,02555066
A1-313	H	4,28395062	5,83216783	0,44135802	0,71333333	1,31717734	14,72727273	7,2	0,00308642	0,86666667	0,06666667	1,04545455	0,5	0,0154321
A2-313	T	4,08668076	6,0625	0,3551797	0,78857143	1,34398529	12,95890411	6,47945205	0,00422833	0,79452055	0,09589041	1	0,63888889	0,01268499
A1-317	H	4,56263736	6,32124352	0,42417582	0,700	1,30551779	13,38235294	7,45901639	0,01538462	0,95081967	0,03278689	0,79411765	0,61290323	0,00659341
A2-317	T	4,72496662	6,71940299	0,44726302	0,77142857	1,39373588	14,98	7,56565657	0,00400534	1,08080808	0,08080808	0,98	0,58333333	0,02803738

E. Programkode

Noen programfunksjoner som jeg har skrevet selv, og som dermed ikke har eksterne kilder.

E1. Cohens *d*

Funksjonen regner ut Cohens *d* for to utvalg. Merk at det er flere varianter av definisjoner for Cohens *d*, blant annet med hensyn til om formelen benytter utvalgsstørrelsen (N) eller antall frihetsgrader (N – 1). Ulike versjoner av funksjoner som regner ut Cohens *d* vil dermed dessverre gi litt ulike verdier; avvikene vil være størst for små utvalg.

```
cohens.d <- function (s1, s2){
  n1 <- length(s1)
  n2 <- length(s2)
  cohens.s <-
    sqrt(((n1-1)*var(s1,na.rm=T)+(n2-1)*var(s2,na.rm=T)) / (n1+n2))
  return((mean(s1,na.rm=T) - mean(s2,na.rm=T))/cohens.s)
}
```

E2. Cramérs V

Funksjonen regner ut Cramérs V på en krysstabell. Utregningen baserer seg på X^2-verdien fra den innebygde funksjonen for kjikvadrat-testen `chisq.test`.

```
cramersv <- function(tabell){
  kji2 <- chisq.test(tabell,correct=F)$statistic
  N <- sum(tabell)
  k <- min(dim(tabell))
  v <- sqrt(kji2/(N*(k-1)))
  return (v)
  }
```

E3. Logit-transformering

Funksjonen regner ut logit for en vektor med p-verdier og en vektor med p+q-verdier, der p er antall positive observasjoner og q er antall negative observasjoner i potensielle omgivelser.

Prosedyren for behandling av nullverdier er beskrevet i avhandlingen. Der p eller q er 0, settes verdien 1/3 inn, som en simulert høyere sannsynlighet enn for 0 observasjoner og lavere sannsynlighet enn for 1 observasjoner. Der antall potensielle omgivelser er 0, brukes medianverdien fra resten av utvalget for p / (p+q).

```
logit2 <- function (x, y, lite=1/3) {
  x[x==0] <- lite
  x[x==y] <- x[x==y] - lite
  xF <- x/y
  xF[y==0] <- median(xF[is.finite(xF)])
  xlogit <- log(xF/(1-xF))
  return(xlogit)
}
```

E4. Entropi

Funksjonen erstatter 0-verdier med 1/3, tilsvarende logit-funksjonen jeg bruker (E3).

```
entropy <- function(v, E=F) {
  v[v==0] <- 1/3
  v.prob <- v / sum(v)
  y <- -sum(v.prob * log2(v.prob))
  ## Returnerer E i stedet for H
  if (E)  {y <- y/log2(length(v))}
  return(y)
}
```

F. Emneregister

G. Litteraturliste

Aitkin, M., Francis, B., Hinde, J. & Darnell, R. (2009). *Statistical modelling in R*. Oxford: Oxford University Press.

Allwood, J. (1998). Some Frequency based Differences between Spoken and Written Swedish. *Papers from the 16th Scandinavian Conference of Linguistics*, 18-29. Lokalisert på http://www.ling.gu.se/~jens/publications/docs076-100/084.pdf

Arecchi, F. T. (2001). *Complexity and emergence of meaning: toward a semiophysics*. Paper presentert ved Complexity and Emergence, Bergamo, Italy. http://www.ino.it/home/arecchi/SezA/fis405.pdf

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

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.

Baron, N. S. (1998). Letters by phone or speech by other means: the linguistics of email. *Language & Communication, 18*(2), 133-170.

Beach, R. & Friedrich, T. (2006). Response to Writing. I C. A. MacArthur, S. Graham & J. Fitzgerald (Red.), *Handbook of writing research* (s. 222-234). New York: Guilford Press.

Beckman, F. S. (1980). *Mathematical foundations of programming*. Addison-Wesley.

Biber, D. (1986). Spoken and written textual dimensions in English: resolving the contradictory findings. *Language, 62*(2), 384-414.

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

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

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

Borgstrøm, C. H. (1973). *Innføring i sprogvidenskap* (2. reviderte utg.). Oslo: Universitetsforlaget.

Chafe, W. L. (1982). Integration and involvement in speaking, writing and oral literature. I D. Tannen (Red.), *Spoken and Written Language* (s. 35-53). Norwood, NJ: Ablex.

Chafe, W. L. & Danielewicz, J. (1987). Properties of Spoken and Written Language. I R. Horowitz & S. J. Samuels (Red.), *Comprehending Oral and Written Language* (s. 83-113): Academic Press.

Chafe, W. L. & Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology, 16*, 383-407.

Chipere, N. (2009). Individual differences in processing complex grammatical structures. I G. Sampson, D. Gil & P. Trudgill (Red.), *Language complexity as an evolving variable* (s. 178-191). Oxford: Oxford University Press.

Chomsky, N. (1957). *Syntactic structures*. Haag: Mouton.

Collier, R. & Werier, C. (1995). When computer writers compose by hand. *Computers and Composition, 12*(1), 47-59.

Collot, M. & Belmore, N. (1996). Electronic Language: A new variety of English. I S. C. Herring (Red.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives* (s. 13-28). Amsterdam: John Benjamins.

Covington, M. A. & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94-100. doi: http://dx.doi.org/10.1080/09296171003643098

Crawley, M. J. (2005). *Statistics: An introduction using R*: John Wiley.

Crawley, M. J. (2007). *The R book*: John Wiley.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Dahl, Ö. (2009). Testing the assumption of complexity invariance : the case of Elfdalian and Swedish. I G. Sampson, D. Gil & P. Trudgill (Red.), *Language complexity as an evolving variable* (s. 50-63). Oxford: Oxford University Press.

Diderichsen, P. (1974[1946]). *Elementær dansk grammatik* (3. udgave utg.). København: Gyldendal.

Dodge, Y. (2010). *The concise encyclopedia of statistics*. New York: Springer.

Epstein, J. A. (2012). Factors Related to Adolescent Computer Use and Electronic Game Use. [Research article]. *ISRN Public Health, 2012*. doi: 10.5402/2012/795868

Eritsland, A. G. (2004). *Skrivepedagogikk: Teori og metode*. Oslo: Samlaget.

Evensen, L. S. (2003). Kvalitetssikring av læringsutbyttet i norsk skriftlig (KAL-prosjektet): Sammendragsrapport (I. f. s.-o. kommunikasjonsstudier, Trans.): NTNU.

Faarlund, J. T., Lie, S. & Vannebo, K. I. (1997). *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Fairbanks, H. (1944). Studies in language behavior. II. The quantitative differentiation of samples of spoken language. *Psychological Monographs, 56*(2), 19-38.

Faraway, J. J. (2005). *Linear models with R*. Boca Raton: Chapman & Hall/CRC.

Faria, J. C., Grosjean, P., Jelihovschi, E. & Farias, P. S. (2015). Tinn-R Editor: GUI for R Language and Environment (Versjon 4.00.03.05). Lokalisert på http://nbcgib.uesc.br/lec/software/editores/tinn-r/en

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

Fitzgerald, J. (1987). Research on Revision in Writing. *Review of Educational Research, 57*(4).

Fjeld, R. V. (2007). Om utviklinga av norsk radio- og fjernsynsspråk: Fra studio og kateter via dagligstue og kosekrok til "ute-på-byen-sted". *Sprog i Norden, 2007*, 47-59.

Flower, L. & Hayes, J. R. (1991). En teori om skriving som kognitiv prosess (K. M. Thorbjørnsen, Overs.). I E. Bjørkvold & S. Penne (Red.), *Skriveteori* (s. 102-127). [Oslo]: LNU / Cappelen.

Foster, P., Tonkyn, A. & Wigglesworth, G. (2000). Measuring spoken language : a unit for all reasons. *Applied Linguistics, 21*(3), 354-375.

Frazier, L. (1985). Syntactic complexity. I D. Dowty, L. Karttunen & A. Zwicky (Red.), *Natural Language Parsing*. Cambridge: Cambridge University Press.

Friederici, A. D. & Brauer, J. (2009). Syntactic complexity in the brain. I T. Givón & M. Shibayama (Red.), *Syntactic complexity : diachrony, acquisition, neuro-cognition, evolution*. Amsterdam: John Benjamins.

Gammerman, A. & Vovk, V. (1999). Kolmogorov Complexity : Sources, Theory and Applications. *The Computer Journal, 42*(4).

Geizer, R. (1967). Psychogrammatical measures. *Communication quarterly, 15*(3), 31-33.

Gell-Mann, M. (1995). What is complexity? *Complexity, 1*(1), 16-19.

Ghasemi, A. & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology Metabolism, 10*(2), 486-489.

Givón, T. & Shibayama, M. (Red.). (2009). *Syntactic complexity : diachrony, acquistion, neuro-cognition, evolution*. Amsterdam: John Benjamins.

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

Goyvaerts, J. (2009). Powergrep: Just Great Software Co. Ltd. Lokalisert på https://www.powergrep.com/

Gries, S. T. (2009). *Statistics for linguists with R : a practical introduction* (Vol. 208). Berlin: Mouton de Gruyter.

Haas, C. (1989). Does the Medium Make a Difference? Two studies of Writing with Pen and Paper and with Computers. *Human-Computer Interaction, 4*(2), 149-169.

Halliday, M. A. K. (1979). *Differences between spoken and written language: Some implications for literacy teaching.* Paper presented at the 4th Australian Reading Conference, Adelaide.

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): Academic Press.

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

Halliday, M. A. K. (1998). Muntlige og skriftlige måter å mene på. I K. L. Berge, P. Coppock & E. Maagerø (Red.), *Å skape mening med språk* (s. 265-291).

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (Third edition utg.). London: Hodder Arnold.

Harrington, S., Shermis, M. D. & Rollins, A. L. (2000). The influence of word processing on English placement test results. *Computers and Composition, 17*(2), 197-210.

Hawisher, G. E. (1986). The effects of word processing on the revision strategies of college students (Publication no. ED268546). fra ERIC

Holmes, D. I. & Forsyth, R. S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and linguistic computing, 10*(2), 111-127.

Horowitz, M. W. & Berkowitz, A. (1964). Structural advantage of the mechanism of spoken expression as a factor in differences in spoken and written expression. *Perceptual and motor skills, 19*, 619-625.

Horowitz, M. W. & Berkowitz, A. (1967). Listening and reading, speaking and writing: an experimental investigation of differential acquisition and reproduction of memory. *Perceptual and motor skills, 24*, 207-215.

Howell, D. C. (2007). *Statistical methods for psychology* (Sixth edition utg.). Belmont: Thomson Wadsworth.

Hudson, R. (2009). Measuring Maturity. I R. Beard, D. Myhill, J. Riley & M. Nystrand (Red.), *The SAGE handbook of writing development* (s. 349-362). Los Angeles: SAGE.

Hultman, T. G. & Westman, M. (1977). *Gymnasistsvenska* (Vol. 167). Lund: Svensklärarföreningen.

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3: National Council of Teachers of English, Champaign I. L.

Hunt, K. W. (1970). *Syntactic maturity in schoolchildren and adults*. Chicago: University of Chicago Press.

Hård af Segerstad, Y. (2002). *Use and Adaptation of Written Language to the Conditions of Computer-mediated Communication.* Doctor of Philosophy. Göteborg University, Göteborg.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(Suppl. 1).

Jensen, B. U. (2005). *Lingvistisk skriftteori og bokmål.* Cand.Philol (Hovedoppgave). Universitetet i Oslo, Oslo.

Jensen, B. U. & Steien, G. B. (kommer). Stabile og variable intonasjonstrekk i flerspråklige taleres idiolekter i to post-S1-språk.

Johannessen, J. B. ([s.a.]). *Oslo-Bergen-taggeren: en grammatisk tagger for bokmål og nynorsk*. Lokalisert, på http://www.tekstlab.uio.no/obt-ny/

Johannessen, J. B., Hagen, K., Lynum, A. & Nøklestad, A. (2012). OBT+stat: A combined rule-based and statistical tagger. I G. Andersen (Red.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian* (s. 51-66). Amsterdam: John Benjamins.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Lund University, Department of Linguistics and Phonetics, Working Papers, 53*, 61-79.

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs, 56*, 1-15.

Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *Journal of linguistics, 43*(2), 365-392.

Karlsson, F. (2009). Origin and maintenance of clausal embedding complexity. I G. Sampson, D. Gil & P. Trudgill (Red.), *Language complexity as an evolving variable* (s. 192-202). Oxford: Oxford University Press.

Kellogg, R. T. (1994). *The psychology of writing*. Oxford: Oxford University Press.

Kellogg, R. T. (2004). Working memory components in written sentence generation. *American Journal of Psychology, 117*(3), 341-361.

Kellogg, R. T. & Mueller, S. (1993). Performance amplification and process restructuring in computer-based writing. *International Journal of Man-Machine Studies, 39*(1), 33-49.

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

Kulbrandstad, L. I. (2003). *Lesing i utvikling : Teoretiske og didaktiske perspektiver*. Bergen: Fagbokforlaget.

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

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

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

Lovász, L. (1997). Information and complexity : (How to measure them?) *The Emergence of Complexity in Mathematics, Physics, Chemistry and Biology : (Proceedings of the Pontifical Academy of Sciences)* (s. 65-80): Princeton University Press. (Lokalisert på http://www.cs.elte.hu/~lovasz/roma.pdf).

Lowie, W. & Seton, B. (2013). *Essential statistics for applied linguistics*. Basingstoke: Palgrave Macmillan.

MacArthur, C. A. (1999). Overcoming barriers to writing: Computer support for basic writing skills. *Reading & Writing Quarterly, 15*, 169-192.

MacArthur, C. A. (2006). The effects of new technologies on writing and writing processes. I C. A. MacArthur (Red.), *Handbook of writing research* (s. 248-274). New York: Guilford.

Malvern, D. D., Richards, B. J., Chipere, N. & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills: Palgrave MacMillan.

McWhorter, J. (2001). The world's simplest grammars are Creole grammars. *Linguistic typology, 5*(2/3).

McWhorter, J. (2008). Why does a language undress? : Strange cases in Indonesia. I M. Miestamo, K. Sinnemäki & F. Karlsson (Red.), *Language complexity : typology, contact, change* (s. 167-190). Amsterdam: John Benjamins.

Meurer, P. (2012a). ASK - Norsk andrespråkskorpus. Bergen: Uni Computing. Lokalisert på http://clarino.uib.no/ask/ask

Meurer, P. (2012b). Corpuscle: a new corpus management platform for annotated corpora. I G. Andersen (Red.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian* (s. 31-50). Amsterdal: John Benjamins.

Meurer, P. (2017). *Corpuscle*. Lokalisert, på http://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. I M. Miestamo, K. Sinnemäki & F. Karlsson (Red.), *Language complexity : typology, contact, change* (s. 23-42). Amsterdam: John Benjamins.

Miestamo, M. (2009). Implicational hierarchies and grammatical complexity. I G. Sampson, D. Gil & P. Trudgill (Red.), *Language complexity as an evolving variable* (s. 80-97). Oxford: Oxford University Press.

Mikkelsen, P.-O. (2016). *Digital sjibbolett? Testing av andrespråksskriving med penn og PC.* Master i kultur- og språkfagenes didaktikk (Masteravhandling). Høgskolen i Hedmark, Hamar.

Miller, G. A. & Chomsky, N. (1963). Finitary models of language users. I R. D. Luce, R. R. Bush & E. Galanter (Red.), *Handbook of mathematical psychology* (Vol. 2, s. 419-491). New York: Wiley.

Miller, G. A. & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and control, 7*, 292-303.

Monsen, M. (2008). *Kommunikativ funksjonalitet kontra formell korrekthet.* Master (Masteroppgave). Høgskolen i Hedmark, Hamar.

Nichols, J. (2009). Linguistic complexity: a comprehensive definition and survey. I G. Sampson, D. Gil & P. Trudgill (Red.), *Language complexity as an evolving variable* (s. 110-125). Oxford: Oxford University press.

. *NIST/SEMATECH e-Handbook of Statistical Methods*. (2012). Lokalisert, på http://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm

Nyström, C. (2000). *Gymnasisters skrivande. En studie av genre, textstruktur och sammanhang* (Vol. 51). Uppsala: Institutionen för nordiska språk vid Uppsala universitet.

Næs, O. (1965). *Norsk grammatikk : elementære strukturer og syntaks* (2. utgave utg.). Oslo: Fabritius & sønners forlag.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*(2), 217-237.

Partee, B. H., ter Meulen, A. & Wall, R. E. (1993). *Mathematical methods in linguistics*. Dordrecht: Kluwer.

Peña, E. A. & Slate, E. H. (2006). Global Validation of Linear Model Assumptions. *Journal of the American Statistical Association, 101*(473), 341-354. doi: http://dx.doi.org/10.1198/016214505000000637

Peña, E. A. & Slate, E. H. (2014). gvlma: Global Validation of Linear Models Assumptions (Versjon R package version 1.0.0.2). Lokalisert på https://CRAN.R-project.org/package=gvlma

Piolat, A. (1991). Effects of word processing on text revision. *Language and education, 5*(4), 255-272.

Piolat, A., Roussey, J.-Y. & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies, 47*, 565-589.

R Core Team. (2016). R: A language and environment for statistical computing. Wien: R Foundation for Statistical Computing. Lokalisert på http://www.R-project.org/

Rijlaarsdam, G. & van den Bergh, H. (2006). Writing process theory: A functional dynamic approach. I C. A. MacArthur, S. Graham & J. Fitzgerald (Red.), *Handbook of writing research* (s. 41-53). New York: Guilford.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20).

Russell, M. & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3).

Russell, M. & Plati, T. (2001). Effects of Computer Versus Paper Administration of a State-Mandated Writing Assessment. *Teachers College Record*.

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. doi: 10.1093/beheco/ark016

Sampson, G. (2002). *Empirical linguistics*. London: Continuum.

Sampson, G. (2009). A linguistic axiom challenged. I G. Sampson, D. Gil & P. Trudgill (Red.), *Language complexity as an evolving variable*. Oxford: Oxford University Press. (Lokalisert på http://www.grsampson.net/ALac.html).

Sampson, G., Gil, D. & Trudgill, P. (Red.). (2009). *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Sarkar, D. (2008). *lattice : multivariate Data Visualization with R*. Dordrecht: Springer.

Saukkonen, P. (1989). Interpreting textual dimensions through factor analysis. *Glottometrika, 11*, 157-171.

Saukkonen, P. (1993). *Grammatical structures as indicators of textual dimensions.* Paper presented at the XVe Congrès International des Linguistes, Quebec.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(July, October), 379-423, 623-656.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3 and 4), 591-611.

Shen, A. (1999). Discussion on Kolmogorov complexity and statistical analysis. *The Computer Journal, 42*(4).

Sotillo, S. M. (2000). Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning & Technology, 4*(1), 82-119.

Statistisk Sentralbyrå. (2016). *Karakterer ved avsluttet grunnskole, 2016*. Lokalisert, på https://www.ssb.no/utdanning/statistikker/kargrs/aar/2016-10-06

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Cambridge: Blackwell.

SyncRO Soft SRL. (2017). Oxygen XML Editor: SyncRo Soft SRL. Lokalisert på https://www.oxygenxml.com/

Tekstlaboratoriet. (2010). Oslo-korpuset: Universitetet i Oslo.

Tekstlaboratoriet. ([s.a.]). *Oslo-Bergen-taggeren*. Lokalisert 2017, på http://www.tekstlab.uio.no/obt-ny/index.html

Teleman, U., Hellberg, S. & Andersson, E. (1999). *Svenska Akademiens grammatik: satser och meningar* (Vol. 4). Stockholm: Svenska Akademien.

Torrance, M. & Galbraith, D. (2006). The processing demands of writing. I C. A. MacArthur, S. Graham & J. Fitzgerald (Red.), *Handbook of writing research* (s. 67-80). New York: Guilford.

Trudgill, P. (2009). Sociolinguistic typology and complexification. I G. Sampson, D. Gil & P. Trudgill (Red.), *Language complexity as an evolving variable* (s. 98-109). Oxford: Oxford University Press.

Urdan, T. C. (2010). *Statistics in plain English* (3. utg.). New York: Routledge.

Vagle, W. (1990). *Radiospråket - talt eller skrevet?: Syntaktiske og pragmatiske tilnærminger i semiotisk perspektiv*. Oslo: Novus.

Vagle, W. (2005a). Jentene mot røkla: Sammenhengen mellom sensuren i norsk skriftlig og utvalgte bakgrunnsfaktorer, særlig kjønn. I K. L. Berge, L. S. Evensen, F. Hertzberg & W. Vagle (Red.), *Ungdommers skrivekompetanse: Norskeksamen som tekst* (Vol. Bind II, s. 237-274). Oslo: Universitetsforlaget.

Vagle, W. (2005b). Tekstlengde + ordlengdesnitt = kvalitet? Hva kvantitative kriterier forteller om avgangselevenes skriveprestasjoner. I K. L. Berge, L. S. Evensen & W. Vagle (Red.), *Ungdommers skrivekompetanse* (Vol. 2, s. 303-386). (Lokalisert på www.universitetsforlaget.no/skrivekompetanse2).

Vagle, W., Sandvik, M. & Svennevig, J. (1994). *Tekst og Kontekst: en innføring i tekstlingvistikk og pragmatikk*: LNU / Cappelen.

Vinje, E. (1993). *Tekst og tolking: Innføring i litterær analyse*: Ad Notam Gyldendal.

Vinje, F.-E. (1977). *Kompendium i grammatisk analyse* (6. utgave utg.). Oslo: Universitetsforlaget.

Wachal, R. S. & Spreen, O. (1973). Some measures of lexical dciversity in aphasic and normal language performance. *Language & Speech, 16*(2), 169-181.

Western, A. (1921). *Norsk riksmåls-grammatikk : for studerende og lærere* (Faksimileutgave utg.). Kristiania: Aschehoug.

Wikipedia. (2017). *Lempel–Ziv–Welch*. Lokalisert, på https://en.wikipedia.org/wiki/Lempel%E2%80%93Ziv%E2%80%93Welch

World Wide Web Consortium. (2016). *Extensible Markup Language (XML)*. Lokalisert, på https://www.w3.org/XML/

Yngve, V. H. (1961). The depth hypothesis. I R. Jakobson (Red.), *Structure of language and its mathematical aspects*. Providence, Rhode Island: American Mathematical Society. (Lokalisert på http://www.google.com/books?hl=no&lr=&id=ou_zOzU9wEwC&oi=fnd&pg=PA130&dq=yngve+model+hypothesis&ots=GvQcd_WheK&sig=GOnMl6Kz_BwPziQLuOsXpytjxPo#v=onepage&q=yngve%20model%20hypothesis&f=false).

Yngve, V. H. (1998). Clues from the Depth Hypothesis: A Reply to Geoffrey Sampson's Review. [Letter to the editor]. *Computational Linguistics, 24*(4), 633-640.

Yngve, V. H. (Red.). (2006). *Hard-science linguistics*. London: Continuum International Publishing.

Zipf, G. K. (1965 [1935]). *The Psycho-Biology of Language : An Introduction to Dynamic Philology*. Cambridge: MIT press.

Ziv, J. & Lembel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory, IT-23*(3), 337-343.

Östlund-Stjärnegårdh, E. (2002). *Godkänd i svenska? Bedömning och analys av gymnasieelevers texter* (Vol. 57). Uppsala: Institutionen för nordiska språk vid Uppsala universitet.