UDC 811.113.4

Mikołaj Sobkowiak
Adam Mickiewicz University in Poznań, Poland

HVORDAN SKRIVER POLSKE FØRSTEÅRSSTUDERENDE? EN KORPUSUNDERSØGELSE AF SYNTAKTISK KOMPLEKSITET I DANSK SOM FREMMEDSPROG[1]

For citation: Sobkowiak M. Hvordan skriver Polske førsteårsstuderende? En korpusundersøgelse af syntaktisk kompleksitet i dansk som fremmedsprog. *Scandinavian Philology*, 2019, vol. 17, issue 1, pp. 36–54.
https://doi.org/10.21638/11701/spbu21.2019.103

Lingvistisk kompleksitet anses for at være en god indikator for sprogindlæreres præstation og udvikling. I de sidste årtier er lingvistisk, og især syntaktisk, kompleksitet blevet et populært og vigtigt forskningsområde inden for andet- og fremmedsprogstilegnelsen, og diverse kompleksitetsindekser har været anvendt i forskningen som målestok for syntaksen og ordforrådet i L2-tekster. I denne artikel undersøger jeg den syntaktiske kompleksitet i tekster skrevet på dansk af unge polakker. Det undersøgte materiale består af eksamensopgaver skrevet af polske danskstuderende efter studiets første årgang, og de stammer fra forskellige årgange over de sidste 20 år. Tekster skrevet på de forskellige årgange varierer til en vis grad angående både lørner- og opgaverelaterede variabler, og jeg anvender en række kompleksitetsindekser for at undersøge disse variablers spor i de undersøgte teksters syntaks. Jeg fokuserer på forskelle på tværs af lørnergrupper, tekstgenrer og de studerendes køn.

Nøgleord: dansk som fremmedsprog, syntaktisk kompleksitet, korpuslingvistik, lørnerkorpus.

1. INDLEDNING

Lingvistisk kompleksitet kan defineres som "the extent to which language produced in performing a task is elaborate and varied" [Ellis, 2003, p. 140], og den kan analyseres på sprogsystemets diverse ni-

[1] Artiklen er en udvidet udgave af forfatterens oplæg holdt under konferencen MUDS17 i oktober 2018.

veauer (jf. [Boulté, Housen, 2014, p. 43]). Der er i faglitteraturen tale om at skelne mellem absolut og relativ lingvistisk kompleksitet, hvor den førstnævnte type er forbundet med de producerede sproglige enheders grammatiske egenskaber samt deres indbyrdes relationer. Termen *absolut kompleksitet* antyder dog, at der er tale om et objektivt og teorineutralt koncept, og derfor anvender man nogle gange den mindre kontroversielle betegnelse *strukturel kompleksitet* i stedet for (jf. [Berggreen, Sørland, 2016]). Til gengæld associeres relativ kompleksitet, som også er kendt som psykologisk kompleksitet, med de pågældende enheders kognitive sværhedsgrad (jf. [Boulté, Housen, 2014, p. 43]). I det følgende anvendes betegnelsen *kompleksitet* i den absolutte/strukturelle betydning, dvs. uden henvisning til det kognitive.

Resultaterne af flere undersøgelser peger på, at lingvistisk kompleksitet er en alment anvendt og troværdig indikator for lørneres sproglige kompetencer (jf. [Boulté, Housen, 2014; Lahuerta Martínez, 2018]), og den har spillet en vigtig rolle i både L2-forskning og -bedømmelse (jf. [Lu, 2017, p. 494; Kyle, Crossley, 2018, p. 333]). Dette gælder undersøgelser af kompleksitet på forskellige niveauer af sproganalyse (syntaks, ordforråd), men det er syntaktisk kompleksitet, der har fået mest opmærksomhed blandt sprogforskere i de seneste år (jf. [Lu, 2017, p. 496]).

Selv blandt undersøgelser af syntaktisk kompleksitet er der stor variation med hensyn til undersøgelsernes fokus og formål. Mens nogle af længdeundersøgelserne fokuserer på udviklingen i kompleksitet over længere perioder (f.eks. [Berggreen, Sørland, 2016; Kowal, 2016]), er formålet med andre undersøgelser både at måle og analysere kompleksitetens udvikling og lede efter potentielle korrelationer mellem værdierne for kompleksitetsindekser og de undersøgte skriftlige opgavers holistiske vurderinger (f. eks. [Boulté, Housen, 2014; Lahuerta Martínez, 2018]). Derimod har andre, f.eks. Moe [Moe, 2012], undersøgt syntaktisk kompleksitet i forhold til forskellige niveauer i Den europæiske fælles referenceramme for sprog (CEFR), mens Lu [Lu, 2010] og Kyle [Kyle, 2016] både har fokuseret på at måle syntaktisk kompleksitet på forskellige kompetenceniveauer og på at evaluere reliabiliteten af et IT-værktøj skabt til de pågældende formål.

Til trods for de nævnte forskelle har de fleste undersøgelser af strukturel syntaktisk kompleksitet det til fælles, at de anvender en række kompleksitetsindekser. Indekserne illustrerer, hver på sin egen måde, udvalgte aspekter af, hvor sofistikerede og varierede de undersøgte

sproglige enheder er. I de sidste år blev nogle af de "klassiske" indekser kritiseret for at være for generelle (og dermed ikke gode nok) indikatorer for lørnernes sproglige kompetencer. I stedet for foreslog nogle forskere anvendelsen af mere "fine-grained" indekser (jf. [Kyle, 2016; Kyle, Crossley, 2018]). Til trods for det er der mange, som fortsat bruger de mere generelle kompleksitetsindekser med gode resultater (f. eks. [Bulté, Housen, 2018; Kowal, 2016; Lu, 2017; Lahuerta Martínez, 2018]).

Formålet med denne artikel er at præsentere resultaterne af en tværsnitsundersøgelse af syntaktisk kompleksitet i tekster skrevet på dansk af unge polakker. Det undersøgte materiale er eksamensopgaver i praktisk dansk skrevet af polske danskstuderende efter studiets første årgang, og de stammer fra udvalgte årgange i perioden 1996–2016. Der er en vis variation på tværs af årgangene angående både lørner- og opgaverelaterede variabler, og jeg vil anvende en række kompleksitetsindekser for at undersøge, hvorvidt disse variablers spor kan ses i de undersøgte teksters syntaks.

I det følgende vil jeg præsentere mine forskningsspørgsmål samt undersøgelsens design, hvor jeg vil lægge særlig vægt på de undersøgte lørnerdata og opmærkning deraf samt de anvendte kompleksitetsindekser (afsnit 2). I afsnit 3 præsenterer jeg og analyserer undersøgelsens resultater, og til sidst drager jeg konklusioner og kommenterer resultaterne (afsnit 4).

2. UNDERSØGELSENS DESIGN

2.1. Problemformulering

Som nævnt ovenfor er det overordnede formål med denne undersøgelse at analysere syntaktisk kompleksitet i korte tekster skrevet på dansk af unge polakker, som stammer fra 4 forskellige lørnergrupper. Generelt kan indholdet i tekster produceret af lørnere, herunder kompleksitet, ordforrådets diversitet samt sprogets korrekthed, påvirkes af mange lørner- og opgaverelaterede variabler (jf. [Granger, 2008, p. 264]), og derfor må ethvert korpus baseret på lørnerdata bygges på grundlag af "vel gennemtænkte designkriteria" [Tono, 2016, p. 48].

Da der er en vis variation mellem de undersøgte tekster angående lørner- og opgaverelaterede variabler, vil jeg med udgangspunkt i deres potentielle effekt på indholdet forsøge at besvare de følgende forskningsspørgsmål:

1. Er der forskelle angående teksternes syntaktiske kompleksitet på tværs af de undersøgte lørnergrupper, og hvis de findes, kan de forbindes med nogle af de omtalte variabler?
2. Er der forskelle angående syntaktisk kompleksitet i forbindelse med lørnervariablen køn?

Begge spørgsmålenes relevans kan begrundes ud fra tidligere forskningsresultater. Blandt sprogtilegnelsesforskere er der næppe tvivl om, at lørneres præstation og deres sprogs kompleksitet kan afhænge af sådanne lørnerrelaterede faktorer som alder, eksponeringstid, kompetenceniveau og modersmål (jf. f.eks. [Holmen, 1990], i forbindelse med dansk som andetsprog). Derudover peger Polio og Yoon [Polio, Yoon, 2018] på, at tekstens genre kan have indflydelse på sprogets kompleksitet, mens bl.a. Lahuerta Martínez [Lahuerta Martínez, 2018] har identificeret forskelle med hensyn til syntaktisk kompleksitet i tekster skrevet af hhv. drenge og piger.

2.2. Korpusdata

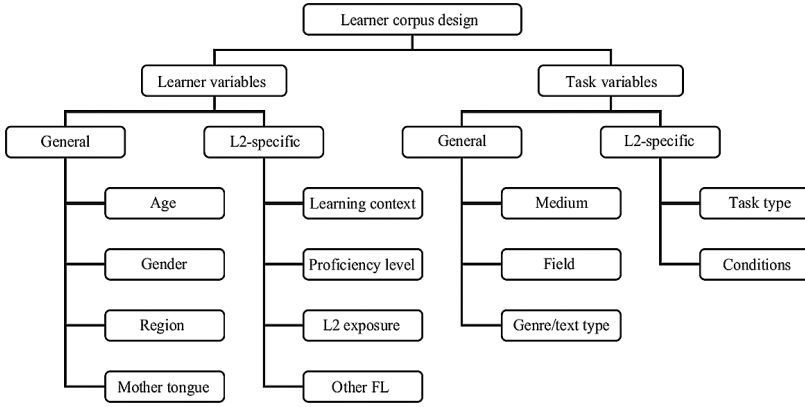
Det undersøgte korpus består af i alt 53 skriftlige eksamensopgaver produceret af 4 grupper polske danskstuderende efter danskstudiets første årgang. I Tabel 1 præsenteres en oversigt over antallet af tekster fra de respektive grupper, herunder kønsfordelingen.

Tabel 1. **Oversigt over tekstantal og kønsfordeling i de 4 undersøgte grupper**

Gruppe	Kvinder	Mænd	I alt
DK1 (n[2])	11	3	14
DK2 (n-1 år)	14	2	16
DK3 (n-10 år)	6	4	10
DK4 (n-20 år)	9	4	13

Selv om man aldrig kan eliminere lørnervariablernes effekt fuldstændigt [Lund, 1997, s. 143], udgør de studerende overordnet set en relativt homogen gruppe med hensyn til Grangers [Grangers, 2008, p. 264] klassifikation

[2] "N" står for det år, teksterne i den pågældende lørnergruppe blev skrevet.



Figur 1. Grangers klassificering af lørner- og opgaverelaterede variabler [Granger, 2008, p. 264]

af generelle lørnervariabler, jf. figur 1. Langt de fleste er 20 år gamle, og de har alle sammen polsk som modersmål. De kommer fra forskellige dele af Polen, men dette synes ikke at have nogen indflydelse på deres tilegnelse af dansk. Der er langt flere kvinder i den undersøgte gruppe end mænd (hhv. 75 % vs. 25 %), hvilket historisk set ikke er nogen markant afvigelse fra det sædvanlige på danskstudiet på det universitet, hvor de studerende har læst.

Angående de L2-specifikke lørnervariabler manifesteres gruppens homogenitet ved, at ingen af de studerende kunne noget dansk ved uddannelsens påbegyndelse. To af grupperne, DK1 og DK2, er naboårgange og har været igennem stort set det samme undervisningsforløb (antal timer, materialer, undervisere, metoder), men der kan forekomme nogle forskelle på det område mellem de to nævnte grupper og DK3 samt DK4. For eksempel har DK4 haft væsentligt flere undervisningstimer end de øvrige grupper, jf. Tabel 2. Forskellene skyldes, at DK3-gruppen skrev deres opgaver 10 år og DK4 20 år før DK1. Af denne grund er det også svært at sige præcist, hvad forskellene angående de anvendte metoder, materialer samt underviserne består i, grundet manglende dokumentation for den ældste gruppe. Ellers har alle de undersøgte studerende i samtlige grupper behersket engelsk og/eller tysk på et højt niveau.

De undersøgte tekster ligner hinanden meget angående de fleste af Granges opgaverelaterede variabler [Granger, 2008, p. 264], da de alle sammen er skrevet i hånden, uden hjælpemidler og under en form for tidspres.

Tabel 2. **Oversigt over de undersøgte lørnergrupper mht. teksternes oprindelsesår og antal undervisningstimer i dansk**

Studieår	År	Antal undervisningstimer
1	n (efter 2010)	270
1	n-1	270
1	n-10	270
1	n-20	420

Emnemæssigt kan de alle sammen siges at være af almensproglig karakter (i modsætning til fagtekster), men de tilhører forskellige genrer, hvilket hænger sammen med de respektive gruppers opgaveformuleringer (jf. Tabel 3).

Tabel 3. **Oversigt over de undersøgte tekster mht. tekstlængde og opgaveformulering**

Gruppe	Gennemsnitlig tekstlængde	Opgaveformulering
DK1	186,64	Skriv en historie ud fra billedet.
DK2	186,94	Skriv en historie ud fra billedet.
DK3	317,50	1. Skriv en historie om hvad der skete før og efter fotoet blev taget. (3 tekster) 2. Danmark og danskere. Hvilket indtryk har du fået af Danmark og danskere i løbet af dette års studier? (7 tekster)
DK4	303,78	? Skriv en historie ud fra billederne.

Som det fremgår af Tabel 3 er der også variation på tværs af grupperne mht. til gennemsnitlig tekstlængde, men dette er ikke relevant for denne undersøgelse, da ingen af de anvendte kompleksitetsmål er forbundet med overordnet tekstlængde som sådan.

2.3. Dataopmærkning og -behandling

Alle teksterne er blevet digitaliseret, og de eneste ændringer foretaget undervejs er, at tal er blevet erstattet af tilsvarende talord og ikkesætningsfinale punktummer blev fjernet. Bortset fra ovennævnte blev

teksternes originale form bibeholdt, inklusive eventuelle ortografiske, syntaktiske og leksikalske afvigelser samt kommateringsfejl.

Herefter blev teksterne opmærket og behandlet i programmet *Essay Tagger*, som er blevet udviklet ved Adam Mickiewicz Universitetet i Poznań specielt til denne undersøgelse. Programmet, som er beslægtet med bl.a. *IA Tagger* (jf. [Jaworski, Jassem, Stroński, 2015]), muliggør semiautomatisk korpusopmærkning på flere niveauer. For denne undersøgelses vedkommende omfattede opmærkningen teksternes opdeling i sætninger, delsætninger[3] og t-enheder (jf. nedenfor).

De undersøgte tekster er i første omgang blevet opdelt i sætninger, hvilket hænger sammen med mit valg af kompleksitetsindekser, som skal bruges i analysen (jf. afsnit 2.4). I modsætning til Berggreen og Sørland [Berggreen, Sørland, 2016] har jeg valgt at opdele teksterne i grafiske sætninger, dvs. at en sætning i min analyse slutter, hvor der står et punktum (som nævnt er ikke-sætningsfinale punktummer, f.eks. i forkortelser, blevet fjernet under tekstbehandling). På denne måde undgik jeg at skulle gætte mig frem til lørnernes intentioner og var i stand til at måle de syntaktiske enheder, de har produceret, og som de opfatter som sætninger.

For at kunne beregne de valgte kompleksitetsindekser, måtte jeg også foretage videre opdelinger (og tilsvarende opmærkning) af samtlige sætninger i teksterne. Jeg skelner i denne forbindelse overordnet set mellem simple og komplekse sætninger, hvor der henholdsvis findes ét og flere bøjede verber / prædikative centre (jf. [Nordborg Nielsen, 2011, s. 355 f.]). Således er (1a) et eksempel på en simpel sætning, mens sætningerne i (1b), (1c) og (1d) er komplekse.

(1) a. Han drikker kaffe.
b. Han drikker kaffe, og hun spiser morgenmad.
c. De spiser morgenmad og drikker kaffe.
d. Han spiser morgenmad, fordi han er sulten.

Med andre ord består en kompleks sætning altså af minimum to delsætninger, mens en simpel sætning kun består af en enkelt delsætning.

Delsætningerne inden for en sætning kan karakteriseres ved forskellige indbyrdes relationer (underordning, sideordning), og af denne grund skelner f.eks. Bulté og Housen [Bulté, Housen, 2014, p. 48] mellem *simple*

[3] Jeg anvender den danske term *delsætning* efter Joel Nordborg Nielsen [Nordborg Nielsen, 2011]. Termens betydning svarer stort set til den engelske betegnelse *clause*.

sentence, *compound sentence*, *complex sentence* og *compound-complex sentence*. Min opdeling er for denne undersøgelses vedkommende meget mindre finkornet, da jeg af hensyn til min analyses formål og design bruger paraplytermen *kompleks sætning* for alle tre af Bulté og Housens kategorier.

Jeg ser dog ikke helt bort fra delsætningernes typer og indbyrdes relationer. En delsætning kan være uafhængig (helsætning) eller den kan være et led i helsætningen (ledsætning). Denne skelnen er relevant inden for lingvistisk kompleksitetsforskning, da en del af de oftest anvendte kompleksitetsindekser er baseret på netop disse forhold og forskelle. I den engelsksprogede litteratur er der i denne forbindelse tale om *T-units* og *clauses*. En *T-unit* (t-enhed) kan defineres som en hovedsætning og eventuelle ledsætninger inden for den samme grafiske sætning, mens en *clause* svarer til en delsætning (jf. f.eks. [Kyle, 2016, p. 10]). En kompleks sætning bestående af to delsætninger kan derfor, afhængigt af delsætningernes indbyrdes relationer, bestå af én eller to t-enheder, jf. hhv. (2a) og (2b).

(2) a. Jens og Jan snakker ofte om de gode gamle dage, når de har drukket vin.
b. Jens og Jan drikker vin, og de snakker om de gode gamle dage.

Mens underordning medfører, at de to delsætninger kun danner en enkelt (kompleks) t-enhed (2a), betragtes to sideordnede delsætninger som to separate t-enheder (2b).

For denne undersøgelse er *Essay Tagger* blevet konfigureret således, at det ud fra den nævnte opmærkning kan beregne de følgende værdier for teksterne: antallet af ord og sætninger, antallet af simple og komplekse sætninger samt antallet af delsætninger og t-enheder. Ud fra disse data beregner programmet de prædefinerede kompleksitetsindekser for en af brugeren defineret gruppe tekster. Indekserne anvendt i denne undersøgelse omtales i afsnit 2.4.

2.4. De anvendte kompleksitetsindekser

I faglitteraturen er der mange syntaktiske kompleksitetsindekser at vælge mellem, og forskellene mellem de enkelte indekser har baggrund i, hvilke niveauer af syntaktisk analyse de undersøgte/målte strukturer stammer fra og hvilke aspekter af strukturernes kompleksitet der skal måles. For eksempel anvendte Bulté og Housen [Bulté, Housen, 2018] et sæt indekser, der dækker såvel sætnings- og t-enhedsniveau som delsætnings- (*clause*-) og fraseniveauet. Lu [Lu, 2010] opdelte derimod sit

sæt af kompleksitetsindekser i fem grupper efter hvilke egenskaber de er designet til at måle. Grupperne omfatter indekser, der vedrører de undersøgte enheders længde, og indekser forbundet med andre forhold som sætningskompleksitet, sideordning, underordning samt bestemte syntaktiske strukturtyper, f.eks. antallet af komplekse nominaler per delsætning.

Tabel 4. **Oversigt over de anvendte kompleksitetsindekser og deres definitioner**

Kompleksitetsindeks	**Engelsk betegnelse**	**Definition**
Type 1: Længdebaserede indekser		
Gennemsnitlig sætningslængde	*Mean length of sentence*	Antal ord / antal sætninger
Gennemsnitlig helsætningslængde	*Mean length of T-unit*	Antal ord / antal helsætninger
Gennemsnitlig delsætningslængde	*Mean length of clause*	Antal ord / antal delsætninger
Type 2: Indekser forbundet med sætningernes overordnede struktur		
Sætningskompleksitet	*Sentence complexity*	Antal delsætninger / antal sætninger
Procenttal for simple sætninger	*Simple sentence ratio*	Antal simple sætninger / antal sætninger
Procenttal for komplekse sætninger	*Complex sentence ratio*[4]	Antal komplekse sætninger / antal sætninger

For denne undersøgelses vedkommende har jeg bestemt mig for at analysere syntaktisk kompleksitet ud fra tre længdebaserede indekser og en enkelt indeks forbundet med sætningernes overordnede sammensætning anvendt af bl.a. Lu. Derudover analyserede jeg, hvor mange af alle sætningerne i de undersøgte tekster der var hhv. simple og komplekse. I Tabel 4 præsenteres

[4] Som nævnt i afsnit 2.3 anvender jeg betegnelsen *kompleks sætning* som paraplyterm for de forskellige typer sætninger bestående af flere and én delsætning nævnt af Bulté og Housen [Bulté, Housen, 2014]. Det samme gælder tilsvarende kompleksitetsindekser — jeg anvender en enkelt indeks angående hvad jeg har valgt at betegne som *komplekse sætninger*, mens Bulté og Housen anvender tre forskellige indekser i stedet for [Bulté, Housen, 2014, p. 48 ff.].

en oversigt over de anvendte indekser samt deres typer, definitioner og tilsvarende engelske betegnelser [jf. Lu, 2010; Bulté, Housen, 2014].

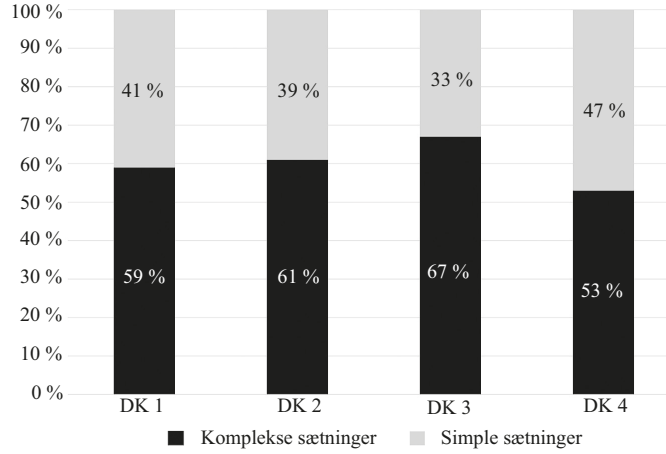
3. RESULTATER

I de følgende afsnit præsenterer og analyserer jeg undersøgelsernes resultater. Først fokuserer jeg på resultaterne på tværs af de undersøgte lørnergrupper (3.1), og derefter ser jeg på forskellene i forbindelse med tekstgenre (3.2) og lørnernes køn (3.3).

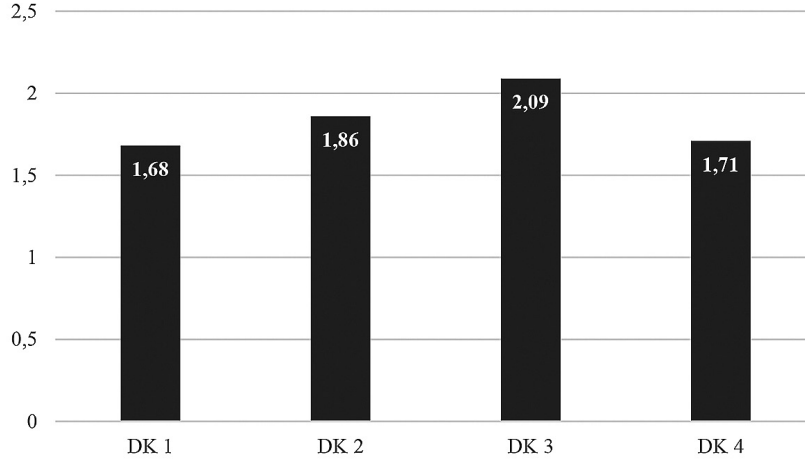
3.1. Resultater på tværs af lørnergrupper

I dette afsnit gennemgår jeg analysens resultater angående de forskellige kompleksitetsindekser med fokus på forskellene mellem de undersøgte årgange.

Det er karakteristisk for polske lørnere på det pågældende niveau at skrive ganske korte og ukomplicerede sætninger, der får teksterne til at lyde staccatoagtige [jf. Sobkowiak, 2017, s. 414]. Alligevel er de fleste sætninger i de undersøgte tekster komplekse, da procenttallene for simple sætninger varierer fra ca. en tredjedel (DK3) til lidt under halvdelen (DK4), jf. Figur 2. Forskellene på tværs af grupperne er dog ikke statistisk signifikante ifølge den gennemførte Mann-Whitney U test/Z-score.



Figur 2. Procenttal for simple og komplekse sætninger i de undersøgte tekster



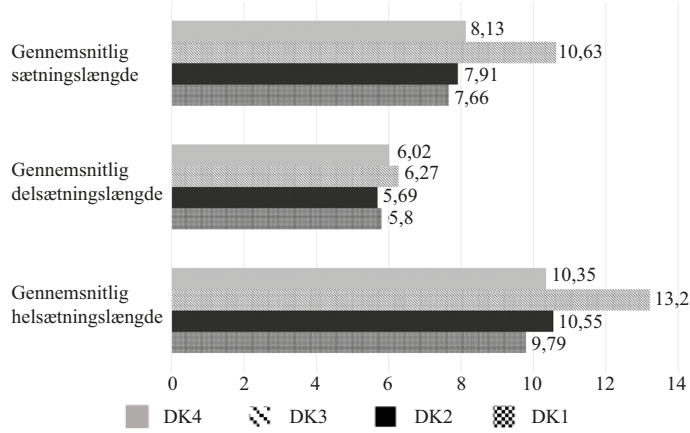
Figur 3. Sætningskompleksitet på tværs af de undersøgte grupper

Angående sætningskompleksitetsindeksen har jeg kunnet konstatere, at sætninger i de undersøgte tekster i gennemsnittet består af ca. 1,8 delsætning. Heriblandt er sætningerne skrevet af DK1-gruppen gennemsnitligt kortest (1,68), mens DK3-gruppen har produceret de længste sætninger og er i øvrigt den eneste af grupperne, hvor gennemsnitstallet ligger over 2 delsætninger pr. sætning, jf. Figur 3.

Derudover er forskellene mellem DK1 og DK3 samt DK3 og DK4 statistisk signifikante ifølge den gennemførte Mann-Whitney U test/Z-score for $p<0.05$ (hhv. $p=0,00652$ og $p=0,02202$).

Der findes også markante forskelle på tværs af grupperne vedrørende de anvendte længdebaserede kompleksitetsindekser, jf. Figur 4. Af selvindlysende grunde er gennemsnitstallene for sætningslængde (mellem 9,79 og 13,22) i det hele taget højere end for hhv. helsætninger (mellem 7,66 og 10,63) og delsætninger (mellem 5,8 og 6,27). Samtidig kan man se, at DK3-gruppen i gennemsnit har produceret de længste sætninger, helsætninger og delsætninger.

Forskellene på tværs af de undersøgte grupper er statistisk signifikante angående gennemsnitlig sætningslængde mellem DK1 og DK3 (Mann-Whitney U test/Z-score for $p<0.05$; $p=0,0151$) samt mellem DK3 og DK4 (Mann-Whitney U test/Z-score for $p<0.05$; $p=0,03753$).



Figur 4. Længdebaserede kompleksitetsindekser i de undersøgte lørnergrupper

Det er de også mellem DK3 og de øvrige grupper angående gennemsnitlig helsætningslængde, jf. Tabel 5.

Tabel 5. **Statistisk signifikans for gennemsnitlig helsætningslængde i de undersøgte grupper (Mann-Whitney U test/Z-score for p<0.05)**

Grupper	p-værdi
DK1 versus DK3	0,00374
DK2 versus DK3	0,00438
DK4 versus DK3	0,02382

Ellers har jeg ikke kunnet konstatere andre statistisk signifikante forskelle mellem grupperne angående de undersøgte længdebaserede kompleksitetsindekser.

DK3 skiller sig tydeligt ud blandt de undersøgte lørnergrupper angående samtlige kompleksitetsindekser. Gruppen har i gennemsnit produceret flest komplekse sætninger i forhold til simple, og samtidig er gruppens sætninger i gennemsnit længste (13,22 ord pr. sætning) og mest komplekse (2,09 delsætninger pr. sætning). Ligeledes er gruppens helsætninger og delsætninger i gennemsnit længere end i tekster produceret af

de øvrige grupper. Forskellene mellem DK3 og samtlige (eller i hvert fald nogle af de) resterende grupper er statistisk signifikante vedrørende sætningskompleksitet samt gennemsnitlig sætnings- og helsætningslængde.

3.2. Resultater på tværs af tekstgenrer

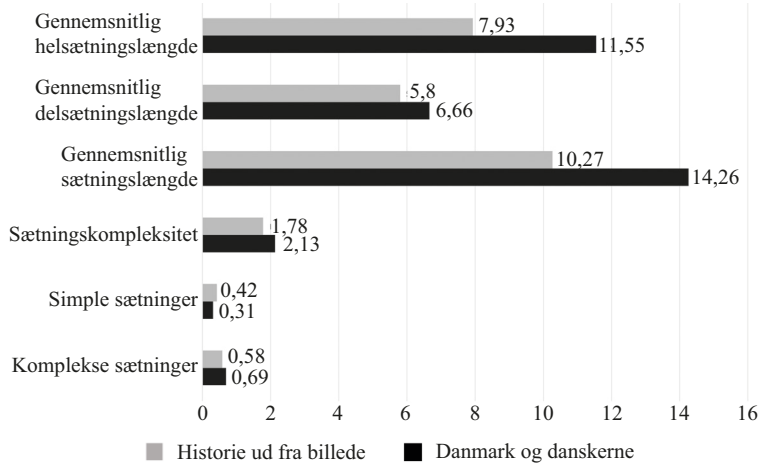
Ud fra ovenstående resultater kan man komme til den konklusion, at DK3 på en eller anden måde kan være forskellig fra de resterende lørnergrupper til trods for, at samtlige studerende udgør en relativt homogen gruppe angående lørnervariablerne (jf. afsnit 2). Dette kan dog være misvisende, for hvad der i virkeligheden potentielt kan ligge til grund for forskellene er, at nogle af teksterne skrevet af DK3-gruppen tilhører en anden genre end de resterende tekster, jf. Tabel 3. Inden for denne gruppe er 30 % af teksterne kreative narrativer ligesom samtlige tekster skrevet af de øvrige grupper. Derimod er 70 % af teksterne skrevet af DK3-gruppen væsentligt forskellige, da de studerende valgte at skrive om, hvad de havde lært om Danmark og danskere siden studiets påbegyndelse. I dette afsnit undersøger jeg, hvordan forskellene er angående de anvendte kompleksitetsindekser på tværs af tekstgenrerne.

Ser man på tallene for de undersøgte kompleksitetsindekser, kan man straks konstatere, at teksterne om Danmark og danskere er mere komplekse end de kreative narrativer, hvor de studerende skulle finde på en historie ud fra et billede, som de havde fået udleveret. Tallene for de respektive kompleksitetsindekser vises i Figur 5.

Forskellene mellem tekster tilhørende de to genrer er ikke kun markante angående absolutte tal, men de er også statistisk signifikante for sætningskompleksitet og samtlige længdemål, jf. Tabel 6.

Tabel 6. **Statistisk signifikans for de undersøgte kompleksitetsindekser på tværs af tekstgenrer (Mann-Whitney U test/Z-score for p<0.05)**

Kompleksitetsindeks	p-værdi
Sætningskompleksitet	0,03236
Gennemsnitlig sætningslængde	0,00544
Gennemsnitlig helsætningslængde	0,00222
Gennemsnitlig delsætningslængde	0,00374



Figur 5. Syntaktisk kompleksitet i de undersøgte tekster på tværs af tekstgenrer

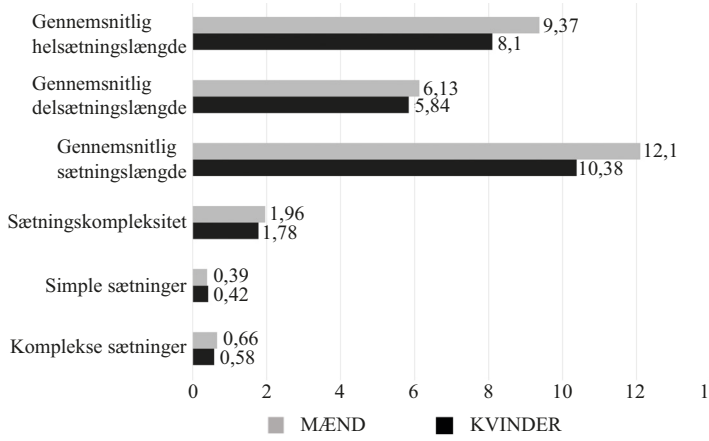
De nævnte kompleksitetsforskelle på tværs af tekstgenrer og ikke mindst deres statistiske signifikans tyder på, at genren kan spille en vigtig (om ikke afgørende) rolle for syntaktisk kompleksitet i tekster skrevet af unge polakker efter danskstudiets første år.

3.3. Resultater på tværs af køn

En del undersøgelser peger på, at den generelle lørnervariabel *køn* kan have indflydelse på, hvor godt man præsterer eller hvor komplekst man udtrykker sig på et fremmedsprog (jf. [Berggreen, Sørland, 2016; Eriksson et al., 2012; Lahuerta Martínez, 2018]). I dette afsnit undersøger jeg, hvorvidt kønsforskellenes spor kan ses i de undersøgte tekster skrevet på dansk af unge polakker.

Som nævnt i Tabel 1 er 40 af de undersøgte tekster skrevet af kvinder og 13 af mænd, hvilket svarer til hhv. 75 % og 25 %. Tallene for de undersøgte kompleksitetsindekser i tekster skrevet af hhv. kvinder og mænd er samlet i Figur 6.

Ser man på ovenstående tal, vil man med det samme lægge mærke til, at tekster skrevet af mænd er mere komplekse, hvilket gælder samtlige undersøgte kompleksitetsindekser. I det undersøgte materiale har mændene nemlig produceret flere komplekse sætninger i forhold til



Figur 6. Syntaktisk kompleksitet i de undersøgte tekster på tværs af køn

simple end kvinderne (hhv. 66 % versus 58 %), og deres sætninger består i gennemsnit af flere delsætninger (1,96 versus 1,78). Ligeledes er såvel de grafiske sætninger som helsætningerne og delsætningerne i gennemsnit længere i mændenes tekster end i kvindernes. Dette er det modsatte af hvad man kunne forvente ud fra tidligere forskningsresultater (jf. f. eks. [Lahuerta Martínez, 2018]). Dog skal man understrege, at ingen af forskellene på tværs af køn er statistisk relevante ifølge den gennemførte Mann-Whitney U test/Z-score for $p<0.05$.

4. KONKLUSIONER

Det overordnede formål med denne artikel har været at præsentere resultaterne af en tværsnitsundersøgelse af syntaktisk kompleksitet i tekster skrevet på dansk af polske danskstuderende. Mere præcist gik opgaven ud på at forsøge at besvare de to forskningsspørgsmål formuleret i afsnit 2.1 og derved (1) prøve at identificere forskellene angående syntaktisk kompleksitet mellem tekster skrevet af de fire undersøgte lørnergrupper samt (2) forskelle angående syntaktisk kompleksitet i forbindelse med lørnervariablen køn.

Som målestok for syntaksen i de undersøgte lørnertekster har jeg brugt 3 længdebaserede kompleksitetsindekser (gennemsnitlig sætnings-, helsætnings- og delsætningslængde) og tre mål forbundet med

sætningernes sammensætning, dvs. sætningskompleksitetsindeksen samt procenttal for komplekse og simple sætninger.

I langt de fleste tilfælde er forskellene mellem lørnergrupperne i forhold til de undersøgte kompleksitetsindekser ikke statistisk signifikante. Dette er til trods for variationen på tværs af grupperne angående antallet af undervisningstimer, undervisningsmetoder og f.eks. adgang til dansksprogede medier uden for undervisningen (jf. Figur 1 og Tabel 2).

Dog er DK3-gruppen forskellig fra de øvrige årgange, idet teksterne skrevet af studerende fra netop denne gruppe er mere komplekse end de resterende tekster, og dette gælder både antallet af komplekse sætninger (i forhold til simple), sætningskompleksitet og samtlige længdebaserede indekser. Derudover er forskellene angående sætningskompleksitet, gennemsnitlig sætnings- og helsætningslængde statistisk signifikante (jf. 3.1).

Resultaterne tyder på, at variablen *tekstgenre* kan have en afgørende eller i hvert fald markant betydning for syntaktisk kompleksitet i polakkernes lørnerdansk på det pågældende niveau. Denne formodning styrkes af dataene i afsnit 3.2, hvor jeg har kunnet konstatere statistisk signifikante forskelle på tværs af tekstgenrer angående samtlige længdebaserede kompleksitetsindekser samt sætningskompleksiteten (jf. Tabel 6).

Mit andet forskningsspørgsmål (jf. 2.1) vedrørte potentielle forskelle i syntaktisk kompleksitet i tekster skrevet af kvinder og mænd. Mod mine egne forventninger støttet af resultaterne fra tidligere forskning viste det sig, at det var teksterne skrevet af mænd, der var mere komplekse end tekster skrevet af kvinder. Forskellene er dog ikke statistisk signifikante og kan skyldes individuelle træk hos de undersøgte mænd. En mere omfattende undersøgelse baseret på et større antal tekster ville være nødvendig for at be- eller afkræfte disse resultater.

Med denne undersøgelse håber jeg at have bidraget til en bedre forståelse af syntaktisk kompleksitet i dansk som fremmedsprog samt til polakkernes tilegnelse af dansk. Jeg er klar over undersøgelsens begrænsninger samt at videre udforskning af emnet er nødvendig for at danne et mere detaljeret billede af samspillet mellem syntaktisk kompleksitet i dansk som fremmedsprog og tekst- og lørnervariablerne. Det samme gælder syntaktisk kompleksitet og dennes rolle i den sprogtilegnelsesproces, polakker gennemgår, når de lærer dansk. Angående videre

forskning synes det hensigtsmæssigt at anvende et større sæt kompleksitetsindekser, som vil dække over flere niveauer af syntaktisk analyse (jf. [Bulté, Housen, 2018]) samt flere diverse aspekter af den syntaktiske kompleksitet (jf. [Lu, 2010]).

Samtidig kan det være til gavn for den videre forsknings resultater at inkludere mere finkornede kompleksitetsindekser (jf. f. eks. [Kyle, 2016]) i analysen og udvide kompleksitetens udforskning ved at analysere lørnernes ordforråd og lørnersprogets korrekthed (som f. eks. [Kowal, 2016]). Den slags data, muligvis også i sammenligning med en analyse af tilsvarende L1-data, har potentiale til betydeligt at bidrage til vores forståelse af, hvordan dansk tilegnes som fremmedsprog af unge polakker.

REFERENCES

Berggreen H. Sørland, K. Syntaktisk kompleksitet i et skriftlig innlærerspråkmateriale. *NOA norsk som andrespråk*, Årgang 32:1–2, 2016. S. 31–75.

Bulté B., Housen A. Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics* 28:1, 2018. S. 147–164.

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

Ellis R. *Task-based language learning and teaching*. Oxford: Oxford University Press, 2003. 398 p.

Eriksson M., Marschik P. B., Tulviste T., Almgren M., Pérez Pereira M., Wehberg S., Marjanovič Umek L., Gayraud F., Kovacevic M., Gallego C. Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology* 30, 2012. P. 326–343.

Granger S. Learner corpora. *Corpus Linguistics. An International Handbook*. Vol. 1. Berlin; de Gruyter, 2008. P. 259–275.

Holmen A. *Udviklingslinier i tilegnelsen af dansk som andetsprog — en kvalitativ, kvantitativ analyse*. Ph.D. thesis, University of Copenhagen. (=*Københavnerstudier i tosprogethed* 12). Kobenhavn: Danmarks Lærerhojskole, 1990. 222 s.

Jaworski R., Jassem K., Stroński K. Manual and Automatic Tagging of Indo-Aryan Languages. *Human Language Technologies as a Challenge for Computer Science and Linguistics*, 2015. P. 550–554.

Kowal, I. *The Dynamics of Complexity, Accuracy and Fluency in Second Language Development*. Kraków: Jagiellonian University Press, 2016. 235 p.

Kyle, K. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Georgia State University, 2016. 186 p.

Kyle K. Crossley S. A. Measuring Syntactic Complexity in L2 Writing Using Fine Grained Clausal and Phrasal Indices. *The Modern Language Journal* 102, 2018. P. 333–349.
Lahuerta Martínez A. C. Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing* 35, 2018. S. 1–11.
Lu X. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing* 34(4), 2017. P. 493–511.
Lu X. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 2010. P. 474–496.
Lund K. *Lærer alle dansk på samme måde? En længdeundersøgelse af voksnes tilegnelse af dansk som andetsprog*. København: Special-pædagogoisk forlag, 1997. 419 s.
Moe E. Syntaktisk kompleksitet og rammeverksnivå. *NORSK PROFIL. Det felles europeiske rammeverket spesifisert for norsk*. Et første steg. Oslo: Novus Forlag, 2012. S. 137–158.
Nordborg Nielsen J. *Russisk Grammatik*. København: Københavns Universitet, Institut for Tværkulturelle og Regionale Studier, Østeuropæisk Afdeling, 2011. 420 s.
Polio C., Yoon H. J. The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics* 28, 2018. P. 165–188.
Sobkowiak M. Om polske danskstuderendes skriftlige præstation. *16. Møde om Udforskningen af Dansk Sprog*. Aarhus: Aarhus Universitet, 2017. S. 405–421.
Tono Y. What is missing in learner corpus design? *Spanish Lørner Corpus Research: Current Trends and Future Perspectives*, Amsterdam; Philadelphia: John Benjamins Publishing Company, 2016. P. 33–52.

Mikołaj Sobkowiak
Adam Mickiewicz University in Poznań, Poland

HOW DO POLISH FIRST-YEAR STUDENTS WRITE? A CORPUS STUDY OF SYNTACTIC COMPLEXITY IN DANISH AS A FOREIGN LANGUAGE

For citation: Sobkowiak M. How do Polish first-year students write? A corpus study of syntactic complexity in Danish as a foreign language. *Scandinavian Philology*, 2019, vol. 17, issue 1, pp. 36–54. https://doi.org/10.21638/11701/spbu21.2019.103

Linguistic complexity is considered a good indicator of language learners' performance and development. In the last few decades, linguistic, and especially syntactic, complexity has become a popular and important field of research within second and foreign language acquisition studies, and various complexity indices have been operationalized in research as a yardstick for the syntax and vocabulary of L2 texts. In this article, I examine the syntactic complexity of texts written in Danish by young Poles. The analyzed material consists of exam papers written by Polish students of Danish phi-

lology after the first year of study, and they come from different learner groups over the last 20 years. There is some variation across the learner groups in terms of both learner- and task-related variables, and I apply a number of complexity indices to examine the traces of these variables in the syntax of the analyzed texts. I focus on differences across learner groups, text genres and the authors' gender.

Keywords: Danish as a foreign language, syntactic complexity, corpus linguistics, learner corpus research.

Mikołaj Sobkowiak
Assistant Professor,
Collegium Novum,
al. Niepodległości 4,
61–874 Poznań, Poland
E-mail: miksobko@amu.edu.pl

Received: March 11, 2019
Accepted: April 22, 2019