Kan AI reproducere fagdisciplinær stemme?

Et komparativt korpusbaseret studie af GPT4's evne til at reproducere fagdisciplinær stemme i AI-genereret sprogvidenskabelig prosa

EA LINDHARDT OVERGAARD & ULF DALVAD BERTHELSEN

ABSTRACT

Formålet med denne artikel er at afdække, i hvilket omfang generative AI-modeller – med GPT4 som eksempel – er i stand til at reproducere fagdisciplinær stemme i dansksproget akademisk prosa. De nye store sprogmodeller kommer med løfter om at forandre vores skrivepraksisser, herunder også akademisk skrivning, men det er stadig uklart, hvad kvaliteten er af de autogenererede bidrag, ikke mindst når modellerne anvendes på mindre sprog som fx dansk. Vi er særligt interesserede i fænomenet fagdisciplinær stemme, fordi det er et relativt velbeskrevet fænomen, der samtidig kan undersøges kvantitativt gennem analyse af korpusteksters overfladestruktur. Vi fokuserer særligt på tre aspekter af fagdisciplinær stemme, henholdsvis stillingtagen, engagement og fagspecifikt ordforråd og undersøger dette kvantitativt gennem en korpusbaseret komparativ undersøgelse, hvor vi sammenligner et korpus bestående af dansksprogede sprogvidenskabelige artikler med et korpus af AI-genereret akademisk prosa med sprogvidenskabeligt indhold. Analysen viser, at de AI-genererede tekster på nogle områder afviger signifikant fra de autentiske sprogvidenskabelige tekster. For kategorien fagspecifikt ordforråd er forskellene relativt store, og for kategorierne stillingtagen og engagement er forskellene relativt små. I de to sidstnævnte kategorier er forskellene så små, at vi med en vis rimelighed kan sige, at de AI-genererede tekster på disse områder reproducerer fænomenet disciplinær stemme på en måde, der fra et kvantitativt perspektiv er svært at skelne fra det, vi ser i de autentiske tekster.

EMNEORD: generativ kunstig intelligens; chatbots; tekstkvalitet; fagdisciplinær stemme; korpuslingvistik

DOI 10.7146/nys.v1i65.143044

1 INDLEDNING

De nye store AI-baserede sprogmodeller som fx Open AI's GPT4 og Googles PaLM kommer med løfter om at forandre såvel vores private som vores professionelle skrivepraksisser, herunder akademisk skrivning. Hvis løfterne skal indfris, giver det sig selv, at det output, modellerne genererer, må være af høj sproglig kvalitet. Modellerne er imidlertid stadig nye, og det er endnu uklart, hvad kvaliteten er af de autogenererede tekster, ikke mindst når modellerne anvendes på mindre sprog som fx dansk (Eke 2023, Ghosh & Caliskan 2023). Der er derfor god grund til at interessere sig for, hvordan disse modeller præsterer, når de skal generere dansksproget output. Det er dog ingen triviel opgave at vurdere modellernes tekstkvalitet. Uanset om man arbejder inden for den Halliday-inspirerede SFL-tradition (Halliday 1985, Halliday & Hasan 1976) eller inden for den danske pragmatisk orienterede grammatiktradition (Hansen & Heltoft 2019, Harder 2006, Togeby 2003, Togeby 2014), er det en central pointe, at de sproglige valg, en afsender træffer, i vid udstrækning træffes relativt til kommunikationssituationen, herunder såvel den konkrete kommunikationssituation som den bredere kulturelle kontekst, hvori kommunikationssituationen er indlejret. Den konkrete sproglige realisering af tekster forventes altså at variere, som en konsekvens heraf, med hensyn til fx ordforråd, tekstlængde, syntaktisk kompleksitet og stilleje. De store sprogmodeller er probabilistiske modeller baseret på den såkaldte transformerarkitektur (Vaswani m.fl. 2017). Dette vil lidt forenklet sige, at det output, de genererer, er den mest sandsynlige ordkombination relativt til prompt og træningsdata, samtidig med at modellerne på tekstniveau er i stand til at forholde sig til både den globale og lokale kontekst. Fordi modellerne er trænet på meget store mængder af naturligt sprog, må vi forvente, at den genre- og registervariation, vi finder i det naturlige sprog, afspejles i de data, modellerne er trænet på. Som følge heraf må vi også forvente, at denne variation i et eller andet omfang reproduceres i det autogenererede output, idet sandsynligheden for, at et givet træk optræder i det autogenererede output, groft sagt er baseret på, hvor ofte og i hvilke kontekster dette træk optræder i modellens træningsdata. Kvaliteten af outputtet kan derfor ikke undersøges generisk, men må nødvendigvis undersøges relativt til kontekst og genre. Sagt på en anden måde kan

vi ikke nøjes med at stille det generelle spørgsmål: Er GPT4 god til at skrive dansk? Vi må i stedet spørge mere specifikt, om GPT4 fx er god til at skrive jobansøgninger på dansk, eller om den er god til at skrive festtaler på dansk.

Som et bidrag til at undersøge kvaliteten af autogenereret sprog fokuserer vi i denne artikel på akademisk skrivning. Vi bruger sprogvidenskabelige tekster som eksempel og fokuserer særligt på fænomenet *fagdisciplinær stemme* (Hyland 2005, 2008). Tekstkvalitet er naturligvis et komplekst fænomen, men vi har valgt at fokusere på fagdisciplinær stemme, fordi det er et relativt velbeskrevet fænomen, der samtidig kan undersøges kvantitativt gennem analyse af korpusteksters overfladestruktur (Fløttum m.fl. 2006, Guinda 2012). Formålet med undersøgelsen er at afdække, i hvilket omfang generative AI-modeller, med GPT4 som eksempel, er i stand til at reproducere fagdisciplinær stemme i dansksproget sprogvidenskabelig prosa. I artiklen præsenterer vi en kvantitativ korpusbaseret komparativ undersøgelse, hvor vi sammenligner et korpus bestående af autentiske danskprogede sprogvidenskabelige artikler med et korpus af AI-genereret akademisk prosa med sprogvidenskabeligt indhold. Korpusset af autentiske sprogvidenskabelige artikler består af 306 artikler, og det AI-genererede korpus består af 300 automatisk genererede artikelfragmenter. Vi fokuserer særligt på to aspekter af fagdisciplinær stemme, henholdsvis stillingtagen og engagement, samt på anvendelsen af fagspecifikt ordforråd, og fænomenerne operationaliseres som en række sproglige træk, herunder bl.a. brugen af pronomener, modalitet og adverbielle udtryk. I forbindelse med sammenligningen benyttes permutationstest for at fastslå, om der er signifikant forskel i forekomsten af de forskellige typer af stemmemarkører mellem de to korpusser.

Artiklen indledes med en præsentation af undersøgelsens teoretiske afsæt samt en redegørelse for operationaliseringen af fagdisciplinær stemme. Herefter følger et metodeafsnit, hvor vi redegør for opbygningen af de to korpusser samt rationalet bag valget af statistiske test. Dette følges af en gennemgang af analysens resultater, og artiklen afsluttes med en diskussion af resultaternes relevans og rækkevidde.

2 BAGGRUND

De seneste år er spørgsmålet om kvaliteten af AI-genereret tekst og anvendelsen af generativ kunstig intelligens blevet stadig mere vigtigt, idet generative AI-værktøjer baseret på store sprogmodeller nu er så veludviklede, at de kan bruges som både kreative sparringspartnere og som værktøjer i forskningsprojekter og endda som medforfattere på videnskabelige tekster. Den videnskabelige debat vedrørende denne udvikling handler især om AI-modellernes evne til at producere sammenhængende og meningsfulde tekster på tværs af forskellige domæner samt dels etiske problemstillinger, dels problematikker relateret til inddragelsen af kunstig intelligens som en aktiv medspiller i forskningsverdenen.

I forskningslitteraturen er debatten fortrinsvis fagdisciplinært funderet, idet brugen af kunstig intelligens oftest undersøges i relation til specifikke videnskabelige domæner. Særligt det sundhedsvidenskabelige område (Clusmann m.fl. 2023, Peng m.fl. 2023), litteraturforskning med fokus på kreativ skrivning (Gunser m.fl. 2022, Hitsuwari m.fl. 2023, Jensen & Sørhaug 2021, Köbis & Mossink 2021) samt generelle akademiske skriveproblematikker (Anderson m.fl. 2023, Dergaa m.fl. 2023, Hinton & Wagemans 2023) fylder i forskningslitteraturen. Disse studier kredser på forskellig vis om især tre overordnede spørgsmål: 1) Om kvaliteten af AI-genereret tekst er sammenlignelig med tekster skrevet af mennesker, 2) om man kan påvise, at en tekst er skrevet ved brug af kunstig intelligens ved at analysere forskellige tekstuelle og lingvistiske aspekter af teksterne, og 3) om AI-modeller på fornuftig vis kan benyttes både til læring for elever og studerende og som værktøj til forskning og på arbejdsmarkedet. Alle er dog enige om, at AI-generering af tekster kræver forbedringer på en række områder, hvis AI-værktøjer skal kunne inddrages i alle aspekter af akademiske og litterære skriveprocesser. Endelig peges der på, at den manglende kvalitet i AI-genereret tekst på nuværende tidspunkt skyldes, at AI-modellerne ikke er trænet på domænespecifikke områder, hvilket bl.a. påvirker modellernes evne til at efterligne menneskelig kreativitet i kunstneriske genrer som fx lyrik (Jensen & Sørhaug 2021: 17). Desuden peges der på, at det er problematisk, at AI-genereret tekst ikke automatisk faktatjekkes, og at modellerne ikke forholder sig kritisk til det indhold,

teksterne er baseret på (Peng m.fl. 2023, Frye 2022). Endelig kritiseres AI-genereret tekst for at mangle afsenderperspektiver (Frye 2022).

Selvom forskningslitteraturen om anvendelse af generativ kunstig intelligens er voksende, er der stadig kun få komparative studier, der sammenligner kvaliteten af AI-generede tekster med tekster skrevet af mennesker. Der argumenteres imidlertid i stigende grad for nødvendigheden af denne type studier (Gunser m.fl. 2022, Hinton & Wagemans 2023, Hitsuwari m.fl. 2023), og vi forsøger med denne artikel at bidrage til feltet med en undersøgelse af GPT4's evne til at generere dansksproget sprogvidenskabelig prosa.

3 FAGDISCIPLINÆR STEMME

Begrebet *stemme* er ikke entydigt defineret, men anvendes om forskellige fænomener i forskellige videnskabelige sammenhænge. Inden for fonetikken interesserer man sig fx for stemmen som fysiologisk og akustisk fænomen (Grønnum 1998), inden for litteraturteorien beskæftiger men sig med stemme som narrativt fænomen (Chatman 1978, Rimmon-Kenan 1983, Wall 1991), og inden for pædagogikken beskæftiger man sig med stemme som et literacy-fænomen relateret til identitet, demokratisk deltagelse og udvikling af skrivekompetencer (MacBeath 2006, Krogh 2012). I denne artikel trækker vi på den funktionelt orienterede forskning i fagdisciplinær skrivning (se fx Biber 2006, Biber & Gray 2015, Hyland 2000, Hyland 2017), herunder særligt Hyland (2005, 2008), der knytter fænomenet stemme til *stillingtagen* (stance) og *engagement* (engagement). Stemme forstås i denne sammenhæng som noget, der får sit tekstlige udtryk i og med afsenderens anvendelse af kommunikative ressourcer med det formål at repræsentere sig selv, sin position og sin relation til andre, herunder modtageren (Hyland 2008: 20). Eksempler på kommunikative ressourcer, der typisk anvendes med dette formål, kunne være modalverber og attitudeadverbier, der bl.a. anvendes til at etablere forpligtende relationer mellem afsender og modtager (Hansen & Heltoft 2019 kap. VI, § 12-13) og til at angive afsenderens holdning til det fremsatte sagforhold (Berthelsen 2007). I forlængelse heraf opfattes stemme som et fænomen knyttet til register og genre. Anvendelsen af kommunikative ressourcer til at udtrykke stemme er således ikke

først og fremmest et stilistisk udtryk for afsenderens kreativitet eller idiosynkrasier, men derimod udtryk for sproglige valg truffet relativt til formål og kommunikationskontekst. Dette giver anledning til den antagelse, at de mønstre, vi ser i anvendelsen af stemme i fagdisciplinær skrivning, i vid udstrækning er udtryk for, at afsenderen tilstræber en tilnærmelse til normer og praksisser accepteret af de fagdisciplinære fællesskaber (Biber & Conrad 2009, Togeby 2014). Som følge heraf kan fænomenet som allerede nævnt ikke undersøges generisk. Af denne grund fokuserer vi i denne artikel på vores egen metier, nemlig akademisk skrivning. Denne indsnævring er imidlertid ikke tilstrækkelig, idet der er store forskelle på skrivekulturen inden for de forskellige akademiske discipliner (Carter 2007, Hyland 2005). For at indsnævre feltet yderligere har vi derfor valgt at fokusere særligt på det sprogvidenskabelige domæne.

Stemme er det vi med Nordahl-Hansen & Kvernbekk (2020) kan kalde en T-term, dvs. en betegnelse for et teoretisk konstrukt, der ikke kan observeres direkte. Det er som følge heraf nødvendigt at fortage en operationalisering af konstruktet i form af en specificering af, hvilke observerbare træk i teksten, der antages at være udtryk for stemme. I sproget kan stemme udtrykkes på en mangfoldighed af måder, hvilket er en udfordring i relation til operationaliseringen, når man ønsker at undersøge fænomenet kvantitativt. For det første må der være tale om træk ved teksten, der kan identificeres ved hjælp af de NLP-værktøjer (Natural Language Processing), der er til rådighed for et givet sprog. Dette forhold vender vi tilbage til i metodeafsnittet. For det andet må der være tale om træk ved teksten, der entydigt kan identificeres. Som eksempel på denne problematik kan nævnes udtrykket 'sikkert'. Det kan dels optræde som både adjektiv og adverbium, dels anvendes på måder, hvor det er den tekstlige kontekst, der afgør, om det har en forstærkende eller modererende effekt. Sammenlign fx brugen af 'sikkert' i 'hun kommer helt sikkert ikke', hvor 'helt' + 'sikkert' har en forstærkende effekt, med 'det er ikke sikkert, at hun kommer', hvor 'ikke' + 'sikkert' har en modererende effekt. Pointen er, at frekvensen af 'sikkert' i de to korpusser ikke i sig selv siger noget om funktionen.

For at kunne balancere mellem disse to kriterier læner vi os op ad

Hyland (2005). Hyland skelner som allerede nævnt mellem to overordnede dimensioner, henholdsvis stillingtagen og engagement. *Stillingtagen* betegner den holdningsmæssige dimension (attitudinal dimension) repræsenteret ved det, man kan kalde tekstens afsenderorienterede træk, mens *engagement* betegner den relationsmæssige dimension (alignment dimension) repræsenteret ved det, man kan kalde tekstens modtagerorienterede træk. Den holdningsmæssige dimension angår afsenderens arbejde med at positionere sig i forhold til det omtalte, hvorimod den relationsmæssige dimension angår afsenderens arbejde med at positionere sig i forhold til modtageren (Hyland 2005: 176).

Hylands analytiske apparat er omfattende, og det er ikke alle kategorier, der umiddelbart kan scores automatisk. Der er dog nogle, der kan, og af disse har vi udvalgt en række kategorier, der alle er prototypiske eksempler på kommunikative ressourcer, der udtrykker stemme. Som udtryk for stillingtagen har vi valgt at arbejde med fire forskellige kategorier: 1) *førstepersonspronomener*, der eksplicit markerer afsenderens tilstedeværelse i teksten, 2) *attitudeadverbialer*, der markerer afsenderens holdning til det omtalte sagforhold og 3) *karakteriserende adjektiver*, der udtrykker afsenderens vurdering af det omtalte. Herudover foretager vi 4) en *sentimentanalyse*. Sentimentanalysen er ikke en del af Hylands analytiske apparat, men er alligevel inkluderet, fordi den giver en indikation af, om en tekst er overvejende positivt eller negativt ladet. Som udtryk for engagement har vi valgt to forskellige kategorier: 5) *andenpersonspronomener*, der eksplicit indikerer en adressering af modtageren og 6) *modalverber*, der indikerer forskellige grader af engagement og forpligtelse mellem afsender og modtager. Udover disse seks kategorier har vi også valgt at inkludere 7) *klassificerende adjektiver* og 8) *fagspecifikke substantiver*. Disse to mål er ikke direkte knyttet til konstruktet fagdisciplinær stemme, men er alligevel medtaget som supplement til stemmemarkørerne, fordi de kan give en indikation af, i hvilket omfang afsenderen tilstræber en saglig og neutral tone eller en mere subjektiv og holdningsorienteret tone (Baumann & Graves 2010).

4 METODE

I det følgende redegør vi indledningsvis for opbygningen af de to tekstkorpusser, der udgør grundlaget for den sammenlignende analyse, her-

efter for den konkrete operationalisering af konstruktet *fagdisciplinær stemme* og endelig for valget af analysemetoder.

4.1 Opbygning af det autentiske sprogvidenskabelige korpus

Den ene halvdel af datamaterialet består af 306 autentiske sprogvidenskabelige artikler og indeholder i alt 1.116.259 ord. Artiklerne stammer fra konferencerapporterne fra konferencerækken MUDS – Møderne om Udforskningen af Dansk Sprog (www.projekter.au.dk/muds/om-muds). Vi har inkluderet alle dansksprogede artikler inklusive tilhørende abstracts fra årgangene 2000-2020. Udover dansksprogede artikler finder man i konferencerapporten også abstracts uden artikler samt enkelte artikler på svensk eller norsk. Disse er udeladt med henblik på at få et så homogent korpus som muligt. Artiklerne behandler et bredt udsnit af sprogvidenskabelige emner, og de er alle blevet fagfællebedømt inden udgivelse. Vi antager derfor, at dette udvalg, selvom det ikke er repræsentativt i formel forstand, giver et godt afsæt for at undersøge, hvordan fagdisciplinær stemme realiseres i dansksproget sprogvidenskabelig prosa.

4.2 Opbygning af det AI-genererede korpus

Den anden halvdel af datamaterialet er et korpus bestående af 300 AI-genererede tekstfragmenter genereret med OpenAI's store sprogmodel GPT4 (OpenAI 2023). Korpusset indeholder i alt 258.766 ord. I arbejdet med at opbygge det autogenererede korpus har vi forsøgt at balancere mellem på den ene side at generere en tilstrækkelig mængde tekst med henblik på at sikre kvaliteten af de kvantitative analyser, og på den anden side at tage hensyn til, at brugen af store sprogmodeller er relativt ressourcekrævende. For det første har GPT4 en begrænsning på, hvor meget tekst der kan genereres ad gangen, hvilket betyder, at det er relativt tidskrævende at producere mange tekster, og for det andet er brugen af store sprogmodeller meget energikrævende, bl.a. fordi processen er computationelt krævende i forhold til både anvendelsen af processorkraft og lagringsplads (Lauridsen m.fl. 2019, Bender m.fl. 2021). De AI-genererede tekstfragmenter er derfor kortere (3-4 sider per tekstfragment) end de autentiske sprogvidenskabelige artikler, men vi har vurderet, at dette ville være tilstrækkeligt til at sikre kvaliteten af de kvantitative analyser.

4.2.1 De AI-genererede tekstfragmenters struktur og indhold

På nuværende tidspunkt kan GPT4 ikke generere sammenhængende tekst, der i struktur og omfang minder om videnskabelige artikler. Med henblik på at gøre de to korpusser sammenlignelige, såvel strukturelt som indholdsmæssigt, valgte vi derfor at sammensætte hvert tekstfragment af tre selvstændige dele, henholdsvis et abstract, et beskrivende afsnit og et diskuterende afsnit. De enkelte dele blev genereret på baggrund af et sæt sprogvidenskabelige nøgleord, som vi genererede via det sprogvidenskabelige korpus ved hjælp af topic modelling, nærmere bestemt Latent Dirichlet Allocation (LDA) (Blei m.fl. 2003). LDA er en probabilistisk maskinlæringsmodel, der bruges til at afdække, hvilke emner der er indeholdt i en gruppe tekster, og på denne baggrund kategorisere teksterne relativt til, hvor godt emnerne karakteriserer de enkelte tekster. Dette sker ved at se på, hvor ofte ordene optræder sammen, samtidig med at der tages højde for, hvor unikke de enkelte ord er for et givent emne. For eksempel vil et ord som "ord" forekomme på tværs af mange tekster i vores korpus, da vi arbejder med sprogvidenskabelige artikler, og dette ord er derfor ikke emnespecifikt. Omvendt er et ord som "dialekt" knyttet til et specifikt område, og ordet vil være tæt relateret med en række andre ord inden for samme område. For at kunne afgøre, hvor mange emner det sprogvidenskabelige korpus skulle inddeles i, evaluerede vi den såkaldte coherence-score for forskellige antal emner (Weston m.fl. 2023). Coherence-scoren fortæller os groft sagt, hvor meningsfulde emnerne er. Analysen viste, at vi fik færrest overlap, hvis vi valgte at inddele teksterne i seks emner. Som resultat af LDA-analysen får vi således seks emner karakteriseret ved en række særligt frekvente nøgleord i lemmaform. Disse nøgleord har vi brugt som afsæt for at prompte GPT4 til at generere tekster. Nøgleordene for de seks emner er følgende:

```
Emne 1 = varietet, sproglig, dansk, dialekt, undersøgelse,
         kort, resultat, rigsdansk, forskellig, respondent
Emne 2 = sætning, subjekt, dansk, flytning, passiv, ledsætning,
         kasus, eksempel, rolle, flytte
Emne 3 = betydning, dansk, moderne, tekst, eksempel, dialo-
         gisk, proposition, nydansk, partikel, udtrykke
```

```
Emne 4 = dansk, sprog, tale, eksempel, sætning, forskellig,
         tekst, form, spørgsmål, sted
Emne 5 = orddannelse, nederlandsk, kvindenavn, participial,
         deverbael, amager-hollandsk, forstå-signaler, fri-
         sisk, deverbal, amager-register
Emne 6 = navn, fornavn, lemma, stavemåde, dansk, form, antal,
         forskellig, sjælden, nyfødt
```

4.2.2 Prompting

I AI-sammenhæng er en prompt en instruks skrevet i naturligt sprog (i modsætning til i computerkode), der fortæller den generative model, hvilket output man ønsker. I instruksen til GPT4 har vi specificeret 1) at teksten skal bestå af henholdsvis et abstract, et redegørende afsnit og et diskuterende afsnit, 2) hvilken teoretisk ramme teksten skal skrives ud fra og 3) en specifikation af de ti nøgleord. Instruksen sendes herefter til OpenAI's API, hvorefter GPT4-modellen returnerer tre separate tekstsekvenser i form af et abstract, en redegørende tekstsekvens og en diskuterende tekstsekvens. Prompten for et tekstfragment ser ud på følgende måde. For hver sekvens specificeres systemets rolle, herefter det ønskede output:

```
input _ 1 = [{"role": "system", "content": "sprogforsker der skri-
ver sprogvidenskabelige artikler på dansk"}, {"role": "user",
"content": "skriv et videnskabeligt abstract på dansk om syn-
taktisk flytning, den teoretiske ramme er generativ grammatik,
keywords: sætning, subjekt, dansk, flytning, 'passiv, ledsæt-
ning, kasus', eksempel, rolle, flytte"}]
```

```
input _ 2 = [{"role": "system", "content": "sprogforsker der skri-
ver sprogvidenskabelige artikler på dansk"}, role": "user",
"content": "skriv et videnskabeligt teoriafsnit om syntaktisk
flytning, den teoretiske ramme er generativ grammatik, keywords:
sætning, subjekt, dansk, flytning, passiv, ledsætning, kasus,
eksempel, rolle, flytte, præsenter relevante teoretiske begre-
ber"}]
```

```
input _ 3 = [{"role": "system", "content": "sprogforsker der skriver sprogvidenskabelige artikler på dansk"}, {"role": "user", "content": "skriv en videnskabelig diskussion om syntaktisk flytning, den teoretiske ramme er generativ grammatik, keywords: sætning, subjekt, dansk, flytning, passiv, ledsætning, kasus, eksempel, rolle, flytte, diskuter fordelene ved relevante teoretiske positioner"}]
```

Med henblik på at opbygge et korpus af passende størrelse blev modellen bedt om at generere 50 tekstfragmenter per instruks. Fordi der er tale om en probabilistisk model, genereres der som følge heraf 50 forskellige tekstfragmenter baseret på samme nøgleord. Når anmodningen om at generere et tekstoutput sendes til API'en, kan man specificere en såkaldt temperatur, hvilket er en indikation af, hvor konservativt modellen gætter. Jo lavere temperatur, jo mere konservativt. I forbindelsen med denne undersøgelse valgte vi den højst mulige temperatur med henblik på at sikre en vis sproglig variation i det autogenererede korpus.

4.3 Præprocessering

I det kvantitative arbejde med vores tekstkorpus har det været en forudsætning, at teksterne er maskinlæsbare, renset for støj og organiseret hensigtsmæssigt. Vi har derfor præprocesseret de rå data (Ondelli 2018: 143-148). Al præprocessering er foretaget ved hjælp af basisbiblioteker i programmeringssproget Python (Van Rossum & Drake 2009), medmindre andet er nævnt.

De sprogvidenskabelige artikler blev præprocesseret på flere niveauer. Efter indsamlingen af artiklerne, som var tilgængelige i tekstscannede pdf-filer, har vi udtrukket tekstscanningslaget og gemt hver artikel separat som rene tekstfiler (.txt). Tekstfilerne er herefter blevet renset for forstyrrende elementer i tekstscanningslaget, fx sidehoved og -fødder. Vi har efterfølgende fjernet længere citater og transskriptioner. Vi har fjernet disse elementer, da det er svært at skelne mellem stemmemarkører, der er relateret til forfatteren og stemmemarkører benyttet i teksteksempler indeholdt i artiklen, når vi benytter digitale værktøjer i analysen. Korte citater i brødteksten har det ikke været muligt at fjerne. De rensede artikler er efterfølgende samlet i et korpus, hvor metadata som

titel og hvilken udgave af konferencerapporten, artiklen kommer fra, er tilføjet. Hver artikel i korpusset er repræsenteret som én lang streng af tegn. De nødvendige opdelinger i sætninger og tokens foretages først i forbindelse med de enkelte analyser.

De AI-genererede tekster er født maskinlæsbare, idét de er genereret digitalt. Outputtet fra tekstgenereringen er delt i tre dele, et abstract, en redegørende tekst og en diskuterende tekst. Outputtet fra GPT4 er i filformatet JSON (.json), hvor informationer om input og forskellige parametre for instruksen til GPT4 også er indeholdt. Vi har derfor udtrukket selve tekstsekvenserne fra JSON-filerne og derefter gemt dem i txt-format. Ligesom for de sprogvidenskabelige artikler renses de autogenererede tekster for forstyrrende elementer, i dette tilfælde primært for formateringskoden "\n" (linjeskift). Desuden har vi forenet de tre tekstelementer (abstract, redegørende afsnit og diskuterende afsnit) til én samlet streng af tegn. Herefter har vi samlet tekstfragmenterne i et korpus, hvor prompten samt JSON-filen er tilføjet som metadata. Vi har afslutningsvis forenet de to korpusser til ét samlet korpus, der indeholder samtlige 606 tekster. Da vi i analysen benytter NLP-værktøjer, der tager en samlet streng af tegn som input, udgør dette korpus det endelige datasæt og er udgangspunkt for de følgende analyser. En del af analyserne kræver dog yderligere præprocessering, men dette beskrives nærmere i afsnittet for de enkelte analyser, idet fx opdeling i tokens eller sætninger er inkluderet i den NLP-pipeline vi benytter. De forskellige korpusser og det endelige datasæt er konstrueret ved hjælp af Pandas-biblioteket (McKinney 2010) i Python.

4.4 Design af analyse

Vores analyse er frekvensbaseret, hvilket vil sige, at vi for hver tekst tæller en række sproglige træk, der antages at repræsentere konstruktet stemme. Trækkene knytter sig til enten afsenderens *stillingtagen* eller *engagement* eller *anvendelse af fagspecifikt ordforråd*. Inden for stillingtagen ser vi på brugen af førstepersonspronomener, attitudeadverbialer, karakteriserende adjektiver og den emotionelle værdi i teksterne, og trækkene vi knytter til engagement, er andenpersonspronomener og modalverber. Den faglige sprogbrug undersøger vi gennem brugen af klassificerende adjektiver og fagspecifikke substantiver.

For at kunne sammenligne teksterne i de to korpusser har vi udregnet, hvor store andele hver kategori udgør af den samlede mængde ord inden for ordklassen, fx andelen af førstepersonspronomener relativt til det samlede antal pronomener i teksten. Vi har herefter beregnet et gennemsnit for hvert korpus. Vi sammenligner desuden spredningen mellem de to grupper med henblik på at undersøge, i hvilket omfang fordelingen i de to grupper minder om hinanden. Alle analyserne, bortset fra sentimentanalysen, er baseret på en annotering af teksterne foretaget ved hjælp af DaCy (Enevoldsen m.fl. 2021). DaCy er et NLP-værktøj baseret på maskinlæring, der er optimeret til brug på danskprogede tekster. DaCy er valgt, fordi modellen har opnået et state-of-the-art-performanceniveau på flere parametre, blandt andet på POS-tagging (parts of speech) (Enevoldsen m.fl. 2021: 210). DaCy tager de rå tekstdata som input og returnerer komplekse dataobjekter, der indeholder en lang række metadata om teksten, herunder POS-tags og information om tekstens syntaktiske struktur i form af såkaldte *universal dependency tags* (se https://universaldependencies.org), der fx indikerer om et verbum er et hjælpeverbum, og om et nominal er subjekt eller objekt. Fremgangsmåden for de frekvensbaserede analyser gennemgås i følgende afsnit, og metoden for evaluering af analyserne gennemgås efterfølgende. Vi gennemgår metoden for den frekvensbaserede analyse i samme rækkefølge som de analytiske kategorier gennemgås i resultatafsnittet, bortset fra at metoden for kategorierne første- og andenpersonspronomener er slået sammen, ligesom klassificerende og karakteriserende adjektiver er slået sammen under ét afsnit. Disse kategorier er slået sammen her, da vi rent teknisk benytter samme fremgangsmåde i forbindelse med analyserne i de pågældende kategorier.

4.4.1 Første- og andenpersonspronomener

Vi har undersøgt brugen af første- og andenpersonspronomener, fx 'jeg', 'mig', 'vi' og 'du', 'dig', 'I', ved at udtrække en liste for hver tekst, der indeholder samtlige førstepersonspronomener i teksten, og ligeledes udtrække en liste, der indeholder samtlige andenpersonspronomener. Til at måle andelen af de to pronomentyper for hver tekst, har vi udtrukket en liste med samtlige pronomener for hver tekst. Denne liste har vi udtrukket ud fra en opslagsliste indeholdende samtlige pronomener

baseret på DaCy's annotation med POS-tagget "PRON". Ud fra disse lister har vi optalt frekvensen af pronomenerne fra hver gruppe af pronomener og udregnet andelen af førstepersonpronomener og andenpersonspronomener for hver gruppe af tekst. Afslutningsvis har vi beregnet en gennemsnitsandel for hvert af de to korpusser.

4.4.2 Attitudeadverbialer

I forbindelse med undersøgelsen af attitudeadverbialerne har vi dels søgt efter attitudeadverbier, dels søgt på bi- og trigrammer, dvs. ordforbindelser bestående af to eller tre ord som fx *'uden tvivl'* og *'ikke med sikkerhed'*. I analysen af adverbier har vi benyttet os af POS-taggene fra annoteringen med DaCy. Da DaCy ikke kan skelne mellem forskellige typer af adverbier, har vi først udtrukket en liste med de 200 mest frekvente adverbier. På denne liste har vi manuelt opmærket alle de adverbier, der er attitudeadverbier, fx 'formentlig' og 'selvfølgelig'. For at finde attitudeadverbialer har vi fundet bi- og trigrams i teksterne og ved hjælp af en stopordsliste fjernet ikke-relevante resultater såsom 'sproglig variation', 'det er' m.fl. Herefter har vi ud fra lister med de 200 hyppigste bigrams og de 200 hyppigste trigrams manuelt opmærket attitudeadverbialerne. Efterfølgende har vi samlet attitudeadverbier og -adverbialer på en liste og på baggrund af denne liste optalt frekvensen af attitudeadverbialer i hver tekst. Vi har udregnet andelen af attitudeadverbialer i forhold til den samlede mængde adverbier i hver tekst, selvom adverbialerne strækker sig ud over adverbiumkategorien. Dette har ingen betydning for resultaterne, da normaliseringen benyttes til sammenligning på tværs af de to korpusser. Vi har desuden udregnet den gennemsnitlige andel for hvert af de to korpusser.

4.4.3 Karakteriserende og klassificerende adjektiver

I analysen af adjektiverne undersøger vi frekvensen af henholdsvis karakteriserende og klassificerende adjektiver i de to korpusser. Vi benytter termerne karakteriserende og klassificerende adjektiver i overensstemmelse med Christensen & Christensen (2019: 75-76). De karakteriserende adjektiver anvendes til at angive egenskaber ved en referent, mens klassificerende adjektiver angiver, at den beskrevne referent tilhører en særlig klasse eller et særligt område. For at kunne lave en opdeling af

karakteriserende og klassificerende adjektiver har vi kigget på den samlede gruppe af adjektiver i vores tekstkorpus. Vi har på baggrund af POS-taggene fra annoteringen med DaCy udtrukket en liste med de 200 mest frekvente adjektiver i datasættet. På denne liste har vi manuelt markeret de adjektiver, der er karakteriserende, fx 'vigtig', 'god' og 'central', og de adjektiver der er klassificerende, fx 'sproglig', 'grammatisk' og 'dansk'. På denne baggrund har vi beregnet andelen af begge typer adjektiver per tekst relativt til det samlede antal adjektiver. Den gennemsnitlige andel af karakteriserende og klassificerende adjektiver er efterfølgende beregnet for hvert korpus.

4.4.4 Modalverber

Brugen af modalverber er som i de foregående tilfælde beregnet som en andel relativt til det samlede antal verber i teksten, hvorefter der er beregnet et gennemsnit for hvert af de to korpusser. Vi har optalt antallet af modalverber i hver tekst ved at søge dem frem ud fra en manuelt opskrevet liste af modalverber. Herefter har vi holdt dette antal op mod det samlede antal verber i hver tekst, som er optalt med udgangspunkt i annoteringen af teksterne. Det samlede antal verber er fundet ved at summere antallet af verber med POS-taggene 'AUX' eller 'VERB', som er hjælpeverberne og de resterende verber.

4.4.5 Fagspecifikke substantiver

Listerne med fagspecifikke substantiver er udtrukket på samme måde som listerne med henholdsvis attitudeadverbier og karakteriserende adjektiver. Først har vi udtrukket en liste over de 200 mest frekvente substantiver i det samlede korpus. Herefter har vi manuelt opmærket de fagspecifikke substantiver. Fagspecifik fortolkes i denne sammenhæng som substantiver tilhørende det sprogvidenskabelige domæne (fx 'syntaks', 'sætning', 'fonem'). Herefter har vi først beregnet andelen af fagspecifikke substantiver per tekst relativt til tekstens samlede antal substantiver og herefter en gennemsnitlig andel for hvert af de to korpusser.

4.4.6 Sentiment

Sentimentanalyse er en kvantitativ måde at måle teksters emotionelle værdi baseret på en antagelse om, at ord har en emotionel værdi. Til

at lave analysen har vi benyttet Python-biblioteket Sentida (Lauridsen m.fl. 2019). Sentida er et af de få værktøjer, der er designet til at lave sentimentanalyse på dansksprogede tekster. Sentida har en præcision på over 80 %, hvilket er højere end andre danske sentiment-værktøjer, der i øjeblikket er tilgængelige. Sentida virker ved at 1) tage en sammenhængende tekst og dele op i ord, 2) gennemgå hvert ord og tildele ordet en sentimentscore baseret på et leksikon over sentiment-bærende ord, 3) tjekke for faktorer, der kan ændre ved sentimentværdien, fx negationer, udråbstegn eller adverbielle modifikatorer, og 4) returnere den gennemsnitlige sentimentscore, dvs. alle ords sentimentværdi summeret og divideret med antallet af ord, der bidrager til sentimentværdien (Lauridsen m.fl. 2019: 45). Fx tilskrives 'god' en positiv værdi, mens 'dårlig' tilskrives en negativ værdi samtidig med, at det sikres, at 'ikke dårligt' tilskrives en positiv værdi. Ved at returnere gennemsnittet tages der højde for variation i tekstlængderne. Sentimentværdierne ligger mellem -5 og 5 og er kategoriseret som vist i følgende tabel 1 (Lauridsen m.fl. 2019: 41):

TABEL 1. SKALA FOR SENTIMENTVÆRDIER

Very strong negative emotion	Strong negative emotion	Slightly negative emotion	Weak negative emotion	Very weak negative emotion	Neutral emotion	Very weak positive emotion	Weak positive emotion	Slightly positive emotion	Strong positive emotion	Very strong positive emotion
-5	-4	-3	-2	-1	0	1	2	3	4	5

(Lauridsen m.fl. 2019)

4.5 Statistisk analyse

Med henblik på at beskrive eventuelle forskelle i brugen af stemmemarkører mellem de autentiske og de AI-generede tekster sammenligner vi de to korpusser ved dels at lave en deskriptiv statistisk analyse, dels at foretage en permutationstest.

4.5.1 Deskriptiv statistisk analyse

For hver af de otte analysekategorier har vi som ovenfor beskrevet beregnet en gennemsnitlig frekvensværdi for hvert af de to korpusser, henholdsvis det autentiske og det AI-genererede. Derudover har vi for hver kategori også undersøgt spredningen i observationerne ved at beregne

standardafvigelsen. Gennemsnittene indikerer, hvilke værdier vores observationer er centreret omkring, og siger noget om, hvor hyppigt hvert af de sproglige træk benyttes i de to grupper af tekster. Standardafvigelsen indikerer, hvor meget observationerne er spredt ud i forhold til gennemsnittet.

Ved at sammenstille gennemsnit og spredning i analysen får vi både et billede af, om hyppigheden af den observerede størrelse i gruppen af AI-genererede tekster er på niveau med de sprogvidenskabelige artikler, og om afstanden mellem observationerne i tekster over og under gennemsnittene er forskellig imellem de to grupper af tekster. Vi kan altså sige noget om, hvorvidt variationen i brugen af et sprogligt træk ligner den naturlige variation, man finder i de autentiske sprogvidenskabelige artikler, og ligeledes om spredningen centrerer sig om det samme midtpunkt i de to grupper af tekster.

Som afsæt for sammenligningen af de frekvensbaserede resultater i analysen benyttes histogrammer til umiddelbar visualisering af de numeriske resultater. Histogrammerne viser fordelingen af målingerne inden for hver analysekategori. Man kan altså ud fra histogrammerne se, hvor mange tekster, der har en bestemt andel af den givne analysekategori, og hvor stor afstand der er mellem de højeste og laveste andele i hver analysekategori. Ud fra histogrammerne kan vi se hvordan data er fordelt, og om der er skævhed i fordelingerne. Er et histogram pænt symmetrisk omkring den gennemsnitlige værdi, er gruppen tilnærmelsesvis normaltfordelt omkring gennemsnittet, hvorimod skæve histogrammer indikerer, at der ikke er tale om en normalfordeling. Det er interessant at se, om fordelingerne for gruppen af AI-genererede tekster og gruppen af sprogvidenskabelige ligner hinanden, idet sammenligningen giver en indikation af, hvor godt GPT4 kan reproducere den menneskelige brug af det givne sproglige træk, herunder i hvilket omfang variationen på tværs af tekster reproduceres.

4.5.2 Permutationstest

I forlængelse af den deskriptive evaluering af resultaterne foretager vi en række permutationstest. En permutationstest er en nonparametrisk statistisk analysemetode, der bruges til at vurdere, om der er signifikant forskel mellem de observerede teststørrelser (fx gennem-

snit) for to grupper. Da vi ikke har grund til at antage, at vores data er normalfordelt, kan vi ikke anvende parametriske test som fx t-test. Permutationstesten er baseret på sampling og er derfor ikke knyttet til bestemte fordelinger af data. I vores analyser benyttes teststørrelserne *forskel i gennemsnit* og *forskel i standardafvigelse*[1] (Chihara & Hesterberg 2019: 52, 421-422).

Med permutationstesten sammenlignes observationer i to grupper ved at permutere observationerne gentagne gange. Det vil sige, at man samler observationerne fra de to grupper i én gruppe og redistribuerer alle observationerne i et nyt datasæt. Herefter beregnes forskellen mellem de to omorganiserede grupper. Denne proces gentages et stort antal gange (typisk 10^3). Afslutningsvis sammenlignes den faktiske observerede forskel mellem de to grupper med alle de permuterede forskelle, hvilket vil sige, at vi beregner en andel for, hvor mange forskelle der er lige så ekstreme eller mere ekstreme end den faktisk observerede forskel. Denne optælling giver en p-værdi, som vi accepterer ved et signifikansniveau på 0,05 (Chihara & Hesterberg 2019: 50-68)[2].

For hvert af de otte parametre tester vi således, om der er signifikant forskel mellem de to gruppers gennemsnitlige andele og standardafvigelser. Vi tester for lighed mellem grupperne. Det vil sige, at nulhypotesen i begge test er, at der ikke er forskel mellem de to grupper. Hvis resultatet er signifikant, dvs. p-værdien er under vores signifikansniveau på 0,05, betyder det, at nulhypotesen forkastes, eller sagt på en anden måde, at der sandsynligvis er en reel forskel mellem grupperne.

5 RESULTATER

I det følgende gennemgår vi resultaterne af analysen. Indledningsvis kaster vi et eksplorativt blik på histogrammerne med henblik på at få et overblik over datamaterialet. Herefter ser vi nærmere på resultaterne af

1 Denne teststørrelse baserer sig på en F-test, som antager normalfordeling. Da vi kun benytter teststørrelsen og ikke udfører selve F-testen, kan det legitimeres at benytte teststørrelsen i kombination med permutationstesten, selvom vi ikke har nogen antagelse om, at vores data har en bestemt fordeling.

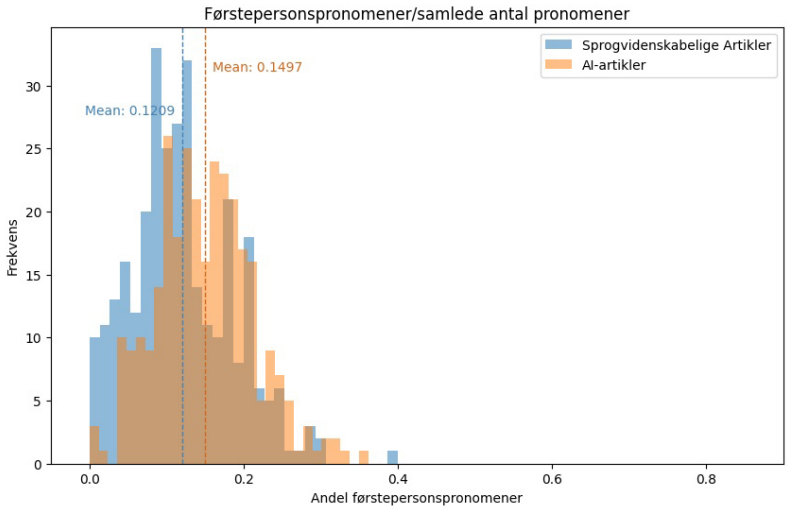
2 For en visuel forklaring af permutationstest, se www.jwilber.me/permutationstest/.

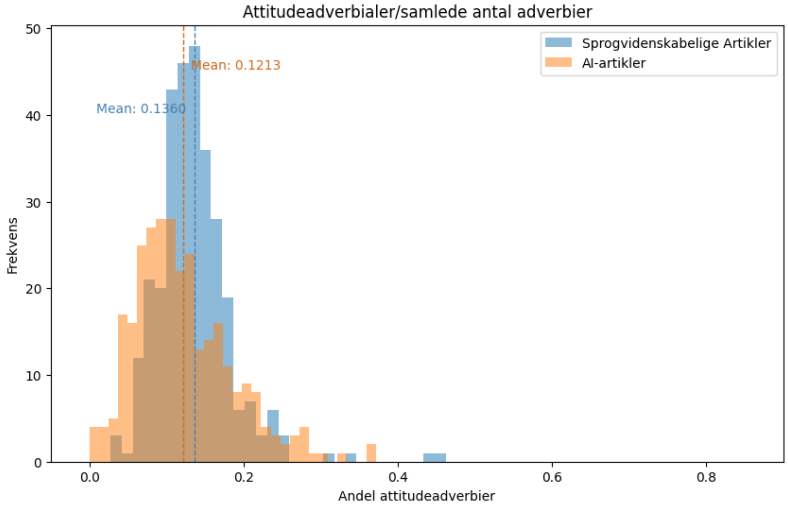
den deskriptive statistiske analyse samt p-værdierne for permutationstestene, og til sidst diskuterer vi resultaterne i relation til fænomenet fagdisciplinær stemme for hver af de tre kategorier *stillingtagen, engagement* og *fagspecifikt ordforråd*.

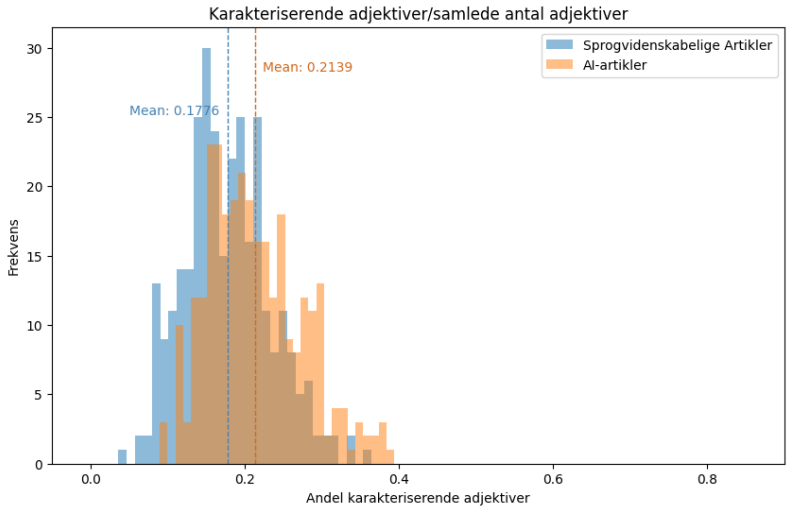
5.1 Histogrammer

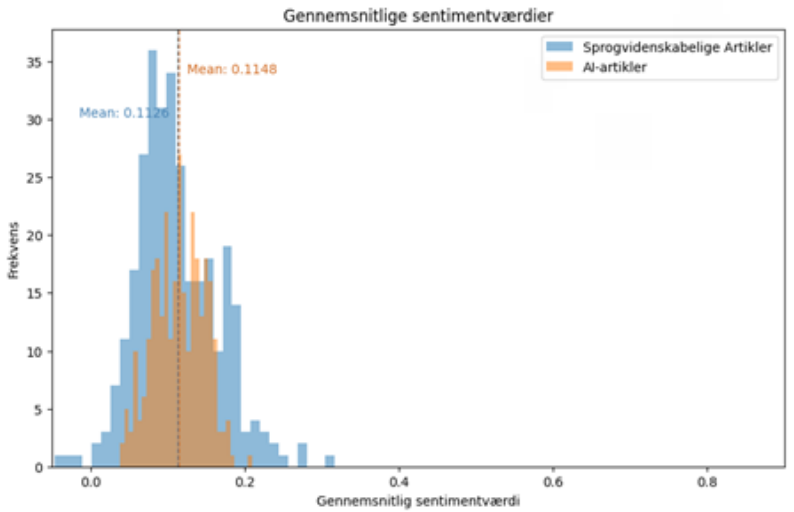
Histogrammer er et godt redskab til at give et visuelt overblik over datamaterialet og en første indikation af, hvad undersøgelsen viser, idet de viser fordelingen af målingerne inden for hver analysekategori. Man kan altså se, hvor mange tekster der har en given andel af de forskellige sproglige træk, og hvor stor afstanden er mellem de højeste og laveste andele i hver analysekategori. Hvert plot indeholder et histogram for gruppen af autentiske sprogvidenskabelige tekster (blåt) og et histogram for gruppen af AI-genererede tekster (orange). De stiplede linjer markerer de to gruppers gennemsnit, og den numeriske værdi er skrevet ud for hver linje, hvilket gør det muligt dels at aflæse, om de to gruppers gennemsnit ligger tæt på hinanden, dels at aflæse i hvilket omfang spredningerne for de to grupper minder om hinanden. X-aksen i histogrammerne viser intervaller med andele inden for den givne analysekategori, og for sentiment-histogrammet viser x-aksen den gennemsnitlige sentimentværdi. Y-aksen viser frekvensen og angiver, hvor mange tekster der falder inden for et givent interval. Hver søjle i histogrammet repræsenterer et interval på x-aksen, og søjlens højde afspejler, hvor mange tekster der falder inden for det givne interval. X-akserne er ensrettet i alle histogrammerne for at lette sammenligningen imellem kategorierne. Y-aksen er derimod justeret i forhold til den enkelte analysekategori for at lette læsningen af de enkelte analysekategorier; en ensretning her ville medføre meget små histogrammer i de fleste kategorier og dermed en mindre detaljeret visualisering.

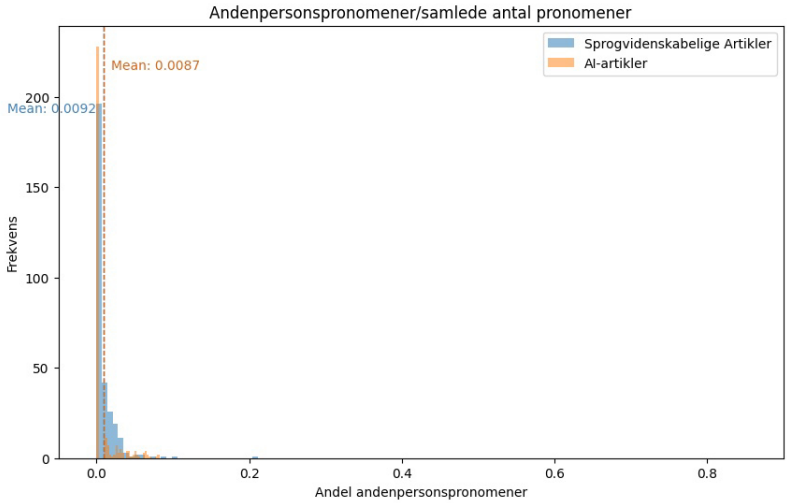
FIGUR 1: HISTOGRAMMER FOR DE OTTE KATEGORIER

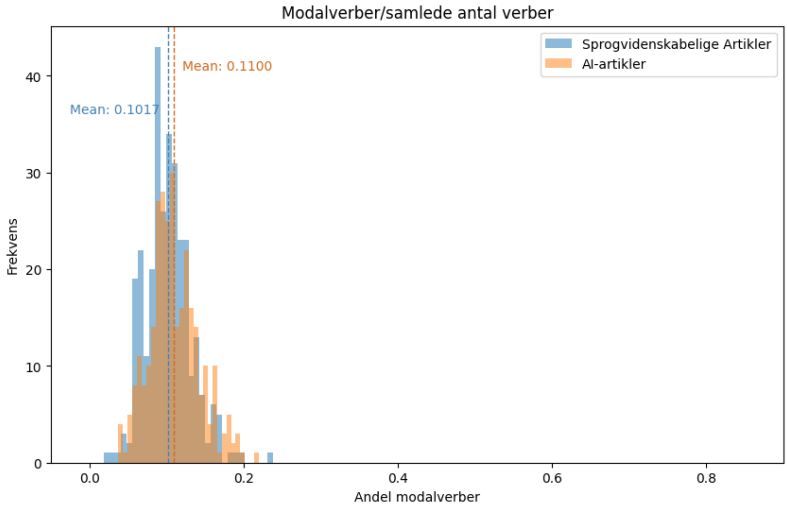


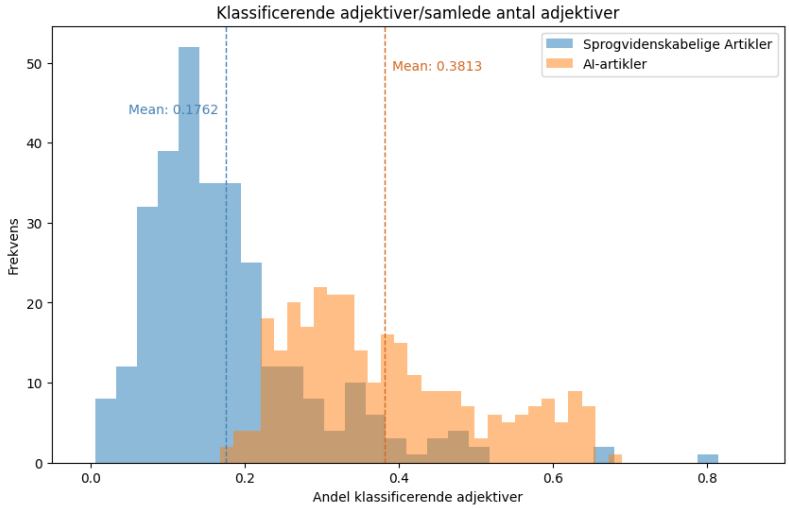


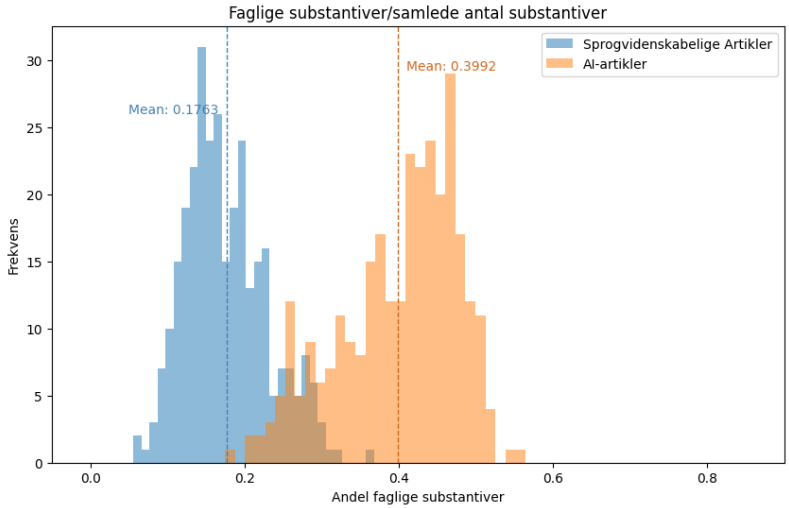

Karakteriserende adjektiver/samlede antal adjektiver
Sprogvidenskabelige Artikler
AI-artikler
Mean: 0.2139
Mean: 0.1776
Frekvens
Andel karakteriserende adjektiver


Gennemsnitlige sentimentværdier
Sprogvidenskabelige Artikler
AI-artikler
Mean: 0.1148
Mean: 0.1126
frekvens
Gennemsnitlig sentimentværdi


Andenpersonspronomener/samlede antal pronomener
Sprogvidenskabelige Artikler
AI-artikler
Mean: 0.0087
Mean: 0.0092
Frekvens
Andel andenpersonspronomener


Modalverber/samlede antal verber
Sprogvidenskabelige Artikler
AI-artikler
Mean: 0.1100
Mean: 0.1017
Frekvens
Andel modalverber


Klassificerende adjektiver/samlede antal adjektiver
Sprogvidenskabelige Artikler
AI-artikler
Mean: 0.1762
Mean: 0.3813
Frekvens
Andel klassificerende adjektiver
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8


Faglige substantiver/samlede antal substantiver
Sprogvidenskabelige Artikler
AI-artikler
Mean: 0.1763
Mean: 0.3992
Frekvens
Andel faglige substantiver
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8

Når vi ser på histogrammerne, er der to ting, der springer i øjnene. For det første ligger gennemsnittene for de autentiske og de AI-genererede tekster i seks ud af de otte kategorier (førstepersonspronomener, attitudeadverbialer, karakteriserende adjektiver, sentimentværdi, andenpersonspronomener og modalverber) relativt tæt på hinanden, og for det andet er der i de samme seks kategorier en meget høj grad af overlap mellem histogrammerne, hvilket indikerer at spredningen er nogenlunde sammenlignelig for de to grupper af tekster i de seks kategorier. De eneste to kategorier, der afviger fra dette mønster, er klassificerende adjektiver og fagspecifikke substantiver. Her ser det til gengæld ud til at være relativt store forskelle både på de to gruppers gennemsnit og på spredningen. De seks første kategorier repræsenterer dimensionerne stillingtagen og engagement, mens de to sidste repræsenterer dimensionen fagspecifikt ordforråd.

Ved første øjekast ser det altså ud til, at GPT4 reproducerer dimensionerne stillingtagen og engagement på en måde, der i høj grad minder om det, vi ser i de autentiske tekster, mens der ser ud til at være en større forskel på brugen af fagspecifikt ordforråd. Her ser det faktisk ud til, at GPT4 i gennemsnit anvender flere fagspecifikke udtryk end det, vi ser i de autentiske tekster. Vi kan imidlertid ikke ud fra histogrammerne afgøre, om der er tale om signifikante forskelle. I det følgende ser vi derfor nærmere på den statistiske analyse af observationerne.

5.2 Statistiske resultater

Tabel 2 viser dels resultaterne af den deskriptive statistiske analyse af de to korpusser, dels resultatet af permutationstesten. Tabellen indeholder de observerede værdier for analysekategorierne førstepersonspronomener, attitudeadverbialer, karakteriserende adjektiver, sentimentværdi, andenpersonspronomener, modalverber, klassificerende adjektiver og fagspecifikke adjektiver, samt en kolonne for henholdsvis den observerede forskel i gennemsnit og den observerede forskel i standardafvigelse for hver kategori. Hver kategori har sin egen række, og de to korpusser er markeret med henholdsvis M (autentiske sprogvidenskabelige artikler) og A (AI-genererede tekster). Værdierne er for alle kategorier angivet i procent, bortset fra sentimentværdien i nederste række, der angiver

den gennemsnitlige sentimentværdi ud fra sentimentskalaen, der som tidligere beskrevet går fra -5 til 5.

TABEL 2: OBSERVEREDE VÆRDIER OG P-VÆRDIER FOR DE TO KORPUSSER

Observator / Kategori	Gruppe	Gennemsnit	Forskel gennemsnit	Standardafvigelse	Forskel standardafvigelse	Minimum	Maksimum
Førstepersonspronomener	M	12,09	2,88***	6,66	0,38	0	40,00
	A	14,97		6,28		0	36,14
Attitudeadverbialer	M	13,60	1,47**	5,01	1,42**	2,73	46,26
	A	12,13		6,43		0	37,14
Karakteriserende adjektiver	M	17,77	3,62***	5,59	0,55	3,56	36,50
	A	21,39		6,14		9,01	39,42
Sentimentværdi	M	0,11	0,0	0,052	0,02***	-0,05	0,35
	A	0,11		0,032		0,04	0,21
Andenpersonspronomener	M	0,92	0,05	1,78	0,04	0	21,11
	A	0,87		1,82		0	8,24
Modalverber	M	10,17	0,83***	3,00	0,23	1,92	23,78
	A	11,00		3,23		3,76	22,02
Klassificerende adjektiver	M	17,62	20,51***	11,25	1,06	0,68	81,37
	A	38,13		12,31		16,83	68,93
Fagspecifikke substantiver	M	17,63	22,29***	5,38	2,35	5,60	36,77
	A	39,92		7,73		17,48	56,41

Signifikansniveau: Markering med * indikerer p <= 0.05, ** indikerer p <= 0.01, og *** indikerer p <= 0.001.

Ser vi på gennemsnittene i tabellen ovenfor, viser det sig, at de AI-genererede tekster har en højere gennemsnitlig andel af førstepersonspronomener, karakteriserende adjektiver, modalverber, klassificerende adjektiver og fagspecifikke substantiver end de autentiske sprogvidenskabelige artikler. For attitudeadverbialerne og andenpersonspronomener forholder det sig omvendt, mens der ikke er nogen forskel på de to gruppers gennemsnitlige sentimentværdi. For andenpersonspronomenerne og modalverberne er forskellen på gennemsnittene på under ét procentpoint, og for førstepersonspronomenerne, attitudeadverbialerne og de karakteriserende adjektiver ligger gennemsnittene mellem 1,47 og 3,62 procentpoint. For både de klassificerende adjektiver og de fagspecifikke substantiver er afstandene mellem de to grupper på over 20 pro-

centpoint, henholdsvis 20,51 for de klassificerende adjektiver og 22,29 for de fagspecifikke substantiver.

Ser vi herefter på forskellen i spredning, dvs. forskellen mellem den observerede standardafvigelse for de to grupper, gælder det for kategorierne førstepersonpronomener og sentimentværdi, at standardafvigelsen for de AI-genererede tekster er lidt lavere end standardafvigelsen for de autentiske tekster. Man finder altså både autentiske tekster med færre og flere førstepersonspronomener, end man ser i gruppen af AI-genererede tekster. Samme billede gør sig gældende for sentimentværdien, hvor der i gruppen af autentiske tekster findes både mere negative, men også mere positive tekster, end man finder i gruppen af AI-genererede tekster. For de øvrige seks kategorier gælder det modsatte. Her er det de AI-genererede tekster, der spreder sig mere omkring gennemsnittet, hvilket vil sige, at man finder AI-genererede tekster med både flere og færre attitudeadverbialer, karakteriserende adjektiver, andenpersonspronomener, modalverber, klassificerende adjektiver og fagspecifikke adjektiver end i gruppen af autentiske tekster. For alle otte kategorier gælder dog, at der er tale om relativt små forskelle i standardafvigelse, der ligger mellem 0,02 (sentimentværdi) som det laveste og 2,35 som det højeste (fagspecifikke substantiver).

Selvom vi som ovenfor beskrevet har obseveret forskelle mellem de to grupper af tekster, er det ikke muligt at vurdere, om forskellene er betydningsfulde alene ud fra sammenligninger af de observerede værdier. Vi har derfor ved brug af permutationstest beregnet p-værdier for de gennemsnitlige forskelle for gennemsnittet og for standardafvigelsen. Resultatet af permutationstestene kan også aflæses i tabel 2, hvor et signifikant resultat er markeret med en eller flere asterisker (* indikerer p <= 0.05, ** indikerer p <= 0.01, og *** indikerer p <= 0.001). P-værdierne for forskel i gennemsnit angiver, om der signifikant forskel imellem de to tekstgruppers gennemsnitlige værdier, og p-værdierne for forskel i standardafvigelse angiver, om der er signifikant forskel i, hvor meget fordelingerne spreder sig omkring gennemsnittet for de to grupper.

Som det fremgår af tabel 2, er der signifikant forskel på gennemsnittene i alle kategorier bortset fra sentimentværdi og andenpersonspronomener, mens der kun er signifikant forskel mellem standardafvigelserne for attitudeadverbialer og sentimentværdi. I de tilfælde, hvor

der er signifikant forskel, er der med stor sandsynlighed en reel forskel på, hvordan de sproglige ressourcer anvendes i de AI-genererede tekster sammenlignet med, hvordan de anvendes i de autentiske sprogvidenskabelige tekster. De signifikante resultater indikerer således, at de målte forskelle mellem tekstgrupperne ikke skyldes tilfældige variationer i teksterne, men at de AI-genererede tekster generelt indeholder en større andel af førstepersonspronomener, karakteriserende adjektiver, modalverber, klassificerende adjektiver og fagspecifikke substantiver og en mindre andel af attitudeadverbialer end de autentiske tekster, samt at spredningen er lidt større i de AI genererede tekster med hensyn til brugen af attitudeadverbialer og lidt større i de autentiske tekster med hensyn til sentimentværdi.

Signifikansresultaterne siger dog ikke i sig selv noget om, hvorvidt der er tale om bemærkelsesværdige eller vigtige forskelle, men peger blot på, hvor der reelt er tale om målbare forskelle. Hvis man som læser vurderede de to grupper af tekster kvalitativt, ville man måske bemærke forskellen på 20 procentpoint i brugen af klassificerende adjektiver og fagspecfikke substantiver, men ville man bemærke en forskel i brugen af førstepersonspronomener, hvor de autentiske sprogvidenskabelige artikler gennemsnitligt indeholder 12,09 % førstepersonspronomener ud af det samlede antal pronomener, mens de AI-genererede tekster gennemsnitligt indeholder 14,97 %? At forskellene er signifikante, indikerer således ikke nødvendigvis, at de er praktisk relevante, men i første omgang blot, at det er værd at se nærmere på den givne kategori for at undersøge, om en eventuel forskel faktisk er betydningsfuld. Dette gør vi i det følgende, hvor vi diskuterer resultaterne for de tre overordnede dimensioner engagement, stillingtagen og fagspecifikt ordforråd.

5.3 Fagdisciplinær stemme

5.3.1 Stillingtagen

Vi begynder med at se nærmere på stillingtagen, dvs. den holdningsmæssige dimension i teksterne relateret til teksternes afsenderorienterede træk, og spørgsmålet er nu, hvad de statistiske resultater siger om GPT4's evne til at reproducere denne dimension. Som vi så ovenfor, var der signifikante forskelle i gennemsnittet for kategorierne førstepersonspronomener, attitudeadverbialer og karakteriserende adjektiver,

men ikke for sentimentværdien. Der var desuden signifikante forskelle på standardafvigelsen i kategorierne attitudeadverbier og sentimentværdi, men ikke i kategorierne førstepersonspronomener og karakteriserende adjektiver. Mere præcist kan vi sige, at GPT4 på den ene side anvender en lidt højere andel af førstepersonspronomener og en lidt højere andel af karakteriserende adjektiver, sammenlignet med de autentiske tekster, mens den på den anden side anvender en lidt lavere andel attitudeadverbialer sammenlignet med de autentiske tekster. Der er dog samtidig lidt større variation i, hvordan GPT4 anvender attitudeadverbialerne, forstået på den måde, at der er AI-genererede tekster, der indeholder både flere og færre attitudeadverbialer sammenlignet med de autentiske tekster. For sentimentværdien er det derimod omvendt. Her er der autentiske tekster, der har både højere og laver sentimentværdi sammenlignet med de AI-genererede.

Den statistiske analyse viser altså, at GPT4 både overforbruger og underforbruger de sproglige ressourcer knyttet til stillingtagen. Dette kunne indikere, at en læser vil møde en tekst, hvor afsenderinstansen på den ene side træder lidt tydeligere frem end i de autentiske tekster, dels gennem eksplicit markering af tilstedeværelse via førstepersonspronomener, fx 'jeg' og 'vi', dels gennem tydeligere markering af subjektive vurderinger via brugen karakteriserende adjektiver, fx 'stor', 'vigtig' og 'relevant', og på den anden side træder lidt mindre tydeligt frem i teksten, idet attitudeadverbialer som fx 'formodentlig', 'måske' og 'helt sikkert', der bl.a. bidrager til tydeliggøre afsenderens holdning til de omtalte, anvendes lidt mindre end det, vi så i de autentiske tekster. Vi så dog samtidig, at de observerede forskelle var små, på maksimalt fire procentpoint, hvilket vil sige, at det er tvivlsomt, om en læser i praksis vil registrere disse forskelle. Der vil naturligvis være tekster, der skiller sig ud med både mere og mindre eksplicit afsendertilstedeværelse, men forskellene er formentlig ikke store nok til, at en læser ved gennemlæsning af flere tekster vil bemærke en generel forskel på, hvordan stillingtagen realiseres i henholdsvis de AI-genererede tekster og de autentiske tekster. Vi hæfter os ved dette, da de små forskelle, til trods for at de i visse tilfælde er signifikante, indikerer, at GPT4 trods alt er tæt på at kunne at reproducere stillingtagen på en måde, der set fra et kvantitativt perspektiv, minder om det, vi ser i de autentiske sprogvidenskabelige tekster.

5.3.2 Engagement

Vi ser herefter på engagement, dvs. den relationsmæssige dimension i teksterne, som er relateret til teksternes modtagerorienterede træk. De sproglige træk, der knytter sig til engagement, er andenpersonspronomener og modalverber, og som vi så ovenfor, var der kun signifikant forskel på den gennemsnitlige andel af modalverber i de to grupper af tekster. Der var ikke signifikant forskel på anvendelsen af andenpersonspronomener, og der var heller ikke signifikante forskelle på spredningen i de to kategorier. Det er altså kun i kategorien modalverber, at GPT4's reproduktion af engagement afviger fra det, vi ser i de autentiske sprogvidenskabelige artikler.

Den statistiske analyse viser her, at GPT4, som vi så med førstepersonspronomener og karakteriserende adjektiver, har et lille overforbrug af modalverber, fx 'kan', 'må' og 'vil'. Dette kunne indikere, at en læser vil møde en tekst, hvor afsenderinstansen, som en konsekvens heraf, fremstår lidt mere engageret i og forpligtet på interaktionen med modtageren sammenlignet med det, vi ser i de autentiske tekster. Som det også var tilfældet med brugen af kommunikative ressourcer knyttet til stillingtagen, er der imidlertid også her tale om en lille forskel, endda en forskel der er endnu mindre, end vi så før, nemlig en observeret forskel imellem de to tekstgrupper på kun 0,83 procentpoint. Resultatet er ganske vist signifikant, men der vil i praksis være tale om en forskel på ganske få ord per tekst, og vi anser det derfor som usandsynligt, at en læser vil bemærke det som en reel kvalitativ forskel på afsenderens engagement i henholdsvis de AI-genererede og de autentiske tekster. Overordnet set indikerer den statistiske analyse altså, idet der kun var signifikant forskel i en ud af de fire kategorier, og idet denne forskel desuden var meget lille, at GPT4 også er i stand til at reproducere engagement på en måde, der, set fra et kvantitativt perspektiv, ligger meget tæt på det, vi ser i de autentiske sprogvidenskabelige artikler.

5.3.3 Fagspecifikt ordforråd

Vi ser herefter på den tredje og sidste dimension, nemlig fagspecifikt ordforråd. De to sproglige træk, der knytter sig til denne dimension, er klasificerende adjektiver og fagspecifikke substantiver. Disse to kategorier er som tidligere nævnt ikke direkte relateret til konstruktet stemme,

men er alligevel interessante, fordi anvendelsen af fagspecifikt ordforråd giver en indikation af, i hvilket omfang afsenderen tilstræber en saglig og neutral tone eller en subjektiv og holdningsorienteret tone. Som vi så ovenfor, var der signifikante forskelle i gennemmsnittene i begge kategorier, men ingen signifikante forskelle i spredningen. Vi så samtidig, at de signifikante gennemsnitlige forskelle for begge kategorier var relativt store, henholdsvis 20,51 for de klassificerende adjektiver og 22,29 for de fagspecifikke substantiver, og at det i begge tilfælde var de AI-genererede artikler, der havde de højeste andele. Den høje andel af fagspecifikke ord i de AI-genererede tekster kan enten indikere, at GPT4 overforbruger fagterminologi, eller at GPT4 bruger fagterminologi i samme grad som i autentiske tekster, men at der i de AI-genererede artikler i mindre grad findes tekst, der ikke indeholder fagterminologi, fx indelednede afsnit eller metatekst, hvilket også vil resultere i en højere andel.

Den statistiske analyse viser således, at der er en reel og relativt stor forskel på, i hvilket omfang de to typer af sproglige ressourcer anvendes i henholdsvis de AI-generede og de autentiske tekster. Det er desuden værd at bemærke, at selv om der ikke er målbare forskelle på spredningen, når vi sammenligner de to grupper af tekster, så er fordelingerne ikke ens. Som de fremgik af plottene i figur 1, så mindede fordelingerne for de øvrige kategorier meget om hinanden. Det var altså ud over få og små målte forskelle også en høj grad af overlap mellem histogrammerne. Dette er ikke tilfældet, når vi se på histogrammerne for de klassificerende adjektiver og de fagspecifikke substantiver. For de klassificerende adjektiver kan vi se, at histogrammet for de autentiske tekster har en tydelig top med mange tekster lidt under gennemsittet samt en hale mod højre med nogle få tekster med meget høje værdier, mens histogrammet for de AI-generede har en mere flad profil med mange tekster spredt ud over et større interval. For de fagspecifikke substantiver kan vi se, at begge histogrammer har en tydelig top samt en lille hale, men at de to histogrammer er spejlvendt, således at de autentiske tekster har en top, der ligger lidt under gennemsittet og en hale, der peger mod højre, mens de AI-genererede har en top, der ligger lidt over gennemsnittet og en hale, der peger mod venstre. Dette er vigtigt, fordi det indikerer, at der er en variation i brugen af klassificerende adjektiver og fagspecifikke substantiver, som ganske vist ikke lader sig måle som signifikant forskel

i standardafvigelse, og som vi ikke umiddlebart kan forklare ud fra vores analyse, men som alligevel ser ud til at være interessant.

Set fra et læserperspektiv er forskellene så store, at det formentlig vil blive bemærket, både fordi det vil være tydeligt for en læser, at der i de AI-genererede tekster i højere grad end i de autentiske tekster anvendes fagspecifikt ordforråd, men også fordi den interne variation i grupperne formentlig er så stor, at det også vil opleves som forskellige skrivestile. Dette kunne fx komme til udtryk ved, at tekster genereret af GPT4, som følge af den noget højere andel af fagspecifikt ordforråd, fx 'syntaks', 'ortografisk' eller 'sociolingvistisk', genrelt vil fremstå mere formelle og mindre subjektive end de autentiske tekster. Samlet set viser den statistiske analyse af fagspecifikt ordforråd altså, modsat hvad vi så, da vi undersøgte de to andre dimensioner, at GPT4 reproducerer denne dimension på en måde, der ligger relativt langt fra den brug af fagspecifikt ordforråd, vi ser i de autentiske tekster.

6 DISKUSSION

Formålet med denne artikel har været at bidrage til at undersøge kvaliteten af AI-genereret akademisk prosa, og vi indledte med at spørge, i hvilket omfang generative AI-modeller, med GPT4 som eksempel, er i stand til at reproducere fagdisciplinær stemme i dansksproget sprogvidenskabelig prosa. Vi har undersøgt dette spørgsmål gennem en kvantitativ komparativ analyse af to korpusser bestående af henholdsvis en samling autentiske dansksprogede sprogvidenskabelige artikler og en samling AI-genererede tekster om sprogvidenskabelige emner. I analysen fokuserede vi på tre dimensioner, dels stillingtagen og engagement, der er direkte knyttet til konstruktet stemme, dels brugen af fagspecifikt ordforråd, der er et vigtigt aspekt af fagdisciplinær skrivning, men ikke specifikt er knyttet til stillingtagen og engagement. Resultatet af analysen var ikke entydigt. På den ene side viste der sig at være signifikante forskelle mellem de AI-genererede og de autentiske tekster i alle de tre overordnede dimensioner. På den anden side var der i flere tilfælde tale om endog meget små forskelle, der formentlig knapt ville blive bemærket af en læser. Groft sagt indikerede analyserne, at stillingtagen i form af en eksplicit tilstedeværende afsenderinstans er lidt tydeligere markeret i de AI-genererede tekster, at afsenderinstansen fremstår mi-

nimalt mere engageret i de AI-genererede tekster, og at de AI-genererede tekster i noget højere grad end de autentiske tekster er præget af fagspecifikt ordforråd.

Som tidligere beskrevet foreligger der kun ganske få studier af kvaliteten af AI-genererede tekster, og der foreligger os bekendt heller ikke benchmarkingresultater for tekstkvalitet af dansksproget akademisk prosa. Det kan derfor på den ene side være vanskeligt umiddelbart at tolke disse lidt tvetydige resultater. Generativ kunstig intelligens er på den anden side et område i voldsom vækst, bl.a. hjulpet på vej af støt stigende mængder af data, og der er derfor ingen særlig grund til at forvente, at GPT4 skulle reproducere fagdisciplinær stemme på en måde, der ligger meget langt fra det, vi ser i de autentiske tekster. Vi vælger derfor at tolke resultaterne som et udtryk for, at GPT4 for kategorierne stillingtagen og engagement, på trods af de målte forskelle, er på vej mod at kunne reproducere fagdisciplinær stemme på en måde, der set fra et kvantitativt perspektiv minder om den, vi ser i autentiske sprogvidenskabelige tekster. I forlængelse heraf er det derfor også overraskende, at den måde, hvorpå GPT4 reproducerede fagspecifikt ordforråd, afveg meget fra det, vi så i de autentiske sprogvidenskabelige tekster, endda på en sådan måde, at GPT4 brugte en væsentlig højere andel af fagspecifikke udtryk end det, vi ser i de autentiske sprogvidenskabelige tekster.

Vi kan ikke umiddelbart forklare denne forskel på baggrund af vores kvantitative undersøgelse. En mulig forklaring kunne være, at MUDS-korpusset ikke er repræsentativt for dansksproget sprogvidenskabelig prosa, og at forskellene ville udjævne sig, hvis korpusset havde været større og bredere. Der kan dog være mange andre årsager, og vores metode muliggør ikke uden videre en undersøgelse af disse årsagssammenhænge. Vores undersøgelse er også begrænset i den henseende, at der udelukkende er tale om en kvantitativ undersøgelse. Den siger således ikke noget om teksternes beskaffenhed, set ud fra et kvalitativt perspektiv, fx om indholdet er faktuelt korrekt, eller om sproget er sammenhængende og meningsbærende.

Vores analyse giver derimod en indikation af, hvordan GPT4-modellen performer på nuværende tidspunkt set fra et kvantitativt perspektiv. Men dette er selvsagt kun et enkelt skridt på vejen mod et

egentligt overblik over kvaliteten af AI-genereret tekst. Både fordi vi kun har undersøgt et enkelt aspekt, nemlig fagdisciplinær stemme, men også fordi det i det hele taget er en stor udfordring at forstå, hvorfor et givet AI-genereret output ser ud, som det gør. Da generativ kunstig intelligens er baseret på probabilistiske modeller trænet på store mængder tekstdata, vil kvaliteten af AI-genereret tekst variere afhængigt af kvaliteten af træningsdatasættet for den generative model og vil i forlængelse heraf også afhænge af, om den genre og det faglige domæne, der genereres tekst inden for, er velrepræsenteret i træningsdatasættet. Kvaliteten af outputtet vil samtidig afhænge af kvaliteten af den statistiske model, der anvendes, og ikke mindst af måden, hvorpå man prompter modellen til at generere tekst.

Denne treenighed, data/model/prompt, gør det til en overordentlig kompleks opgave at vurdere kvaliteten af AI-genereret output, idet enhver ændring i bare én af de tre vil medføre ændringer i outputtet. Mængden af data samt indførelsen af transformerarkitekturen (Vaswani m.fl. 2017) har revolutioneret måden, vi tænker generativ kunstig intelligens på, og der er ingen grund til at tro, at udviklingen ikke vil fortsætte. Hertil kommer, at prompting ikke er en eksakt videnskab. Der er i bogstaveligste forstand uendeligt mange muligheder for at konstruere en prompt, samtidig med at det ikke på nogen måde er transparent, hvordan promptens design påvirker strukturen af outputtet. Dette er et område, der i sig selv kræver meget mere forskning.

Det er derfor ikke tilstrækkeligt at undersøge kvaliteten af autogenereret tekst ud fra hverken et enkelt fagspecifikt perspektiv eller et mere generisk synspunkt. Vi har tværtimod brug for mange flere undersøgelser, der skal undersøge tekstkvaliteten inden for mange forskellige felter, både inden for de akademiske discipliner og inden for andre områder. Det er desuden vigtigt, at der ikke kun er tale om datalogisk orienterede evalueringer af modellernes performance, men at man også ud fra en bred vifte af lingvistiske og kommunikative kriterier vurderer kvaliteten af de nye generative sprogværktøjer på både et mikro- og et makroniveau. Vi betragter, i forlængelse af dette, vores undersøgelse som et indledende skridt i retning af en systematisk og fortløbende indsats for at monitorere kvaliteten af de til enhver tid tilgængelige generative AI-redskaber.

7 AFSLUTTENDE BEMÆRKNINGER

De nye store sprogmodeller kommer som nævnt med løfter om at forandre vores skrivepraksisser, men hvis dette løfte skal indfries, forudsætter det, at kvaliteten af det autogenererede materiale har et rimeligt kvalitetsniveau. Vi har i denne artikel skitseret én mulig måde, hvorpå sprogvidenskabeligt baserede strategier for at teste kvaliteten af sprogværktøjer baseret på generativ kunstig intelligens kan udvikles. Dette er imidlertid ikke den eneste opgave, vi står overfor. Som vores undersøgelse har vist, har vi i høj grad også brug for mere empirisk viden om, hvordan autentiske tekster faktisk ser ud fra et kvantitativt perspektiv. Hvis vi ikke ved det, har vi reelt ikke noget sammenligningsgrundlag for vurderingen af de AI-genererede tekster. Endelig er der den udfordring, at AI-genererede tekster formentlig i mange tilfælde vil optræde uredigerede, fx i forskellige sammenhænge på de sociale medier og i nyhedsmedierne, men at det på sigt kun vil være en del af den AI-genererede tekst, vi vil støde på i vores kommunikative omgang med hinanden. Selvom ideen om at implementere AI-baserede skriveværktøjer i sin skrivepraksis for nogen kan virke fjern, så er realiteten nemlig den, at det allerede benyttes i vid udstrækning. Fx er AI-baserede værktøjer allerede implementeret i skriveprogrammer som Word og Google Docs. Vi vil altså med stor sandsynlighed se, at AI-modellerne vil blive en form for samarbejdspartner inden for det akademiske område, men også inden for mange andre områder. Det er således ikke nok at undersøge kvaliteten af AI-generede tekster i sig selv. Fremover vil det i høj grad også være nødvendigt at undersøge kvaliteten af hybridtekster, dvs. tekster skrevet i et samarbejde mellem mennesker og modeller.

Der er naturligvis al mulig grund til at forholde sig både forsigtigt og kritisk til de nye omkalfatrende, generative teknologier. Der er dog samtidig grund til en vis optimisme i den forstand, at kriterierne for, hvad en god tekst er, ikke kan leveres af de generative teknologier selv. De må derimod komme fra de enkelte faglige domæner, herunder sprog- og tekstvidenskaberne. De generative teknologier har måske nok potentiale til at gøre skrivearbejdet nemmere i kvantitativ forstand, men næppe i kvalitativ forstand, hvis tekstens kompleksitet overstiger den, vi finder i indkøbssedler, bageopskrifter og andre skabelonagtige genrer. At bruge generative teknologier konstruktivt til at skabe kvalitetstekster er med

andre ord ikke noget, der kræver mindre viden om sprog og kommunikation sammenlignet med tidligere skrivepraksisser. Tværtimod kræver produktion af vellykkede hybridtekster indgående kendskab til normer og kvalitetskriterier for forskellige genrer og fremstillingsformer og kalder derfor i høj grad på, at sprog- og tekstvidenskaberne engagerer sig kritisk konstruktivt i arbejdet med at udvikle og kvalitetssikre de nye generative teknologier.

Ea Lindhardt Overgaard, ph.d.-studerende
Institut for Kommunikation og Kultur, Nordisk Sprog og Litteratur
elt@cc.au.dk

Ulf Dalvad Berthelsen, lektor
Institut for Kommunikation og Kultur, Nordisk Sprog og Litteratur
udb@cc.au.dk

LITTERATUR

Anderson, N. m.fl. 2023. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sport — Exercise Medicine* 9(1). 1–4. DOI: 10.1136/bmjsem-2023-001568.

Baumann, J.F. & M.F. Graves. 2010. What Is Academic Vocabulary? *Journal of Adolescent & Adult Literacy*. Hoboken: Blackwell Publishing Ltd. 54(1). 4–12. DOI: 10.1598/JAAL.54.1.1.

Bender, E.M. m.fl. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 610–623. DOI: 10.1145/3442188.3445922.

Berthelsen, U.D. 2007. Attitudeadverbialerne i semantisk og pragmatisk perspektiv. *MUDS - Møderne om Udforskningen af Dansk Sprog* 11. 52–63.

Biber, D. 2006. *University language: a corpus-based study of spoken and written registers*. Philadelphia: Benjamins.

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

Biber, D. & B. Gray. 2015. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511920776.

Blei, D.M., A.Y. Ng & M.I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3. 993–1022.

Carter, M. 2007. Ways of Knowing, Doing, and Writing in the Disciplines. *College composition and communication*. Urbana: National Council of Teachers of English 58(3). 385–418.

Chatman, S. 1978. *Story and discourse: narrative structure in fiction and film*. Ithaca: Cornell University Press.

Chihara, L.M. & T.C. Hesterberg. 2019. *Mathematical Statistics with Resampling and R*. John Wiley & Sons.

Christensen, L.D. & R.Z. Christensen. 2019. *Dansk Grammatik*. Odense: Syddansk Universitetsforlag.

Clusmann, J. m.fl. 2023. The future landscape of large language models in medicine. *Communications Medicine*. Nature Publishing Group 3(1). 1–8. DOI: 10.1038/s43856-023-00370-1.

Dergaa, I. m.fl. 2023. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport* 40(2). 615–622. DOI: 10.5114/biolsport.2023.125623.

Eke, D.O. 2023. ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology* 13. DOI: 10.1016/j.jrt.2023.100060.

Enevoldsen, K., L. Hansen & K. Nielbo. 2021. DaCy: A unified framework for Danish NLP. *Ceur Workshop Proceedings*, vol. 2989, 206–216.

Fløttum, K., T. Dahl & T. Kinn. 2006. *Academic Voices: Across languages and disciplines*. Amsterdam: John Benjamins Publishing Company.

Frye, B.L. 2022. Should Using an AI Text Generator to Produce Academic Writing Be Plagiarism? *Fordham Intellectual Property, Media & Entertainment Law Journal*.

Ghosh, S. & A. Caliskan. 2023. *ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages*. arXiv. DOI: 10.48550/arXiv.2305.10510.

Guinda, C.S. 2012. Proximal Positioning in Students' Graph Commentaries. K. Hyland & C. Sancho Guinda (red.), *Stance and Voice in Written Academic Genres*, 166–183. London: Palgrave MacMillan UK.

Grønnum, N. 1998. *Fonetik og fonologi: almen og dansk*. København: Akademisk Forlag.

Gunser, V.E. m.fl. 2022. The Pure Poet: How Good is the Subjective Credibility and Stylistic Quality of Literary Short Texts Written with an Artificial Intelligence Tool as Compared to Texts Written by Human Authors? *Proceedings of the Annual Meeting of the Cognitive Science Society* 44. 1744–1750.

Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Edward Arnold.

Halliday, M.A.K. & R. Hasan. 1976. *Cohesion in English*. London: Longman.

Hansen, E. & L. Heltoft. 2019. *Grammatik over det danske sprog*. København: Det Danske Sprog- og Litteraturselskab.

Harder, P. 2006. Dansk Funktionel Lingvistik. *NyS – Nydanske Sprogstudier* 34(34–35). 92-130. DOI: 10.7146/nys.v34i34-35.13459.

Hinton, M. & J.H.M. Wagemans. 2023. How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument & Computation*. 14(1). 59–74. DOI: 10.3233/AAC-210026.

Hitsuwari, J. m.fl. 2023. Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior* 139. 107502. DOI: 10.1016/j.chb.2022.107502.

Hyland, K. 2000. *Disciplinary discourses: social interactions in academic writing* (Applied Linguistics and Language Study). Harlow: Longman, an imprint og Pearson Education.

Hyland, K. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*. 7(2). 173–192.

Hyland, K. 2008. Disciplinary voices: Interactions in research writing. *English Text Construction*. 1(1). 5–22. DOI: 10.1075/etc.1.1.03hyl.

Hyland, K. 2017. *The essential Hyland: studies in applied linguistics*. London: Bloomsbury Academic. DOI: 10.5040/9781350037939.

Jensen, C.Z. & E. Sørhaug. 2021. *The Perfect Rap Lyrics - AI Generated Rap Lyrics That Are Better Than Lyrics from Existing Popular and Critically Acclaimed Rap Songs*. NTNU Master thesis. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2834977.

Krogh, E. 2012. Literacy og stemme – et spændingsfelt i modersmålsfaglig skrivning. S. Ongstad (red.), *Nordisk morsmålsdidaktikk: forskning, felt og fag*. Oslo: Novus.

Köbis, N. & L.D. Mossink. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior* 114. 1–12. DOI: 10.1016/j.chb.2020.106553.

Lauridsen, G.A., J.A. Dalsgaard & L.K.B. Svendsen. 2019. SENTIDA: A New Tool for Sentiment Analysis in Danish. *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift* 4(1). 38–53.

MacBeath, J. 2006. Finding a voice, finding self. *Educational Review*. Routledge 58(2). 195–207. DOI: 10.1080/00131910600584140.

McKinney, W. 2010. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference* 56–61. DOI: 10.25080/Majora-92bf1922-00a.

Nordahl-Hansen, A. & T. Kvernbekk. 2020. Construct Validity in Scientific Representation: A Philosophical Tour. *Nordisk tidsskrift for pedagogikk & kritikk* 6. 88–99. DOI: 10.23865/ntpk.v6.1704.

Ondelli, S. 2018. Treat Texts as Data but Remember They Are Made of Words: Compiling and Pre-processing Corpora. A. Tuzzi (red.), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, 133–150. Springer International Publishing. DOI: 10.1007/978-3-319-97064-6_7.

OpenAI. 2023. ChatGPT. https://openai.com/chatgpt. (8 januar, 2024).

Peng, Y. m.fl. 2023. AI-generated text may have a role in evidence-based medicine. *Nature Medicine*. Nature Publishing Group 29(7). 1593–1594. DOI: 10.1038/s41591-023-02366-9.

Rimmon-Kenan, S. 1983. *Narrative fiction: contemporary poetics.* Reprint. London: Methuen.

Togeby, O. 2003. *Fungerer denne sætning?: funktionel dansk sproglære*. København: Gad.

Togeby, O. 2014. *Bland blot genrerne - ikke tekstarterne!: om sprog, tekster og samfund.* Frederiksberg: Samfundslitteratur.

Van Rossum, G. & F.L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vaswani, Ashish, m.fl. 2017. Attention is all you need. *Advances in neural information processing systems* 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wall, B. 1991. *The narrator's voice: the dilemma of children's fiction*. London: MacMillan Press Limited. DOI: 10.1007/978-1-349-21109-8.

Weston, S.J. m.fl. 2023. Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 6(2). 1–13. DOI: 10.1177/25152459231160105.