Nordisk nettverk for edisjonsfilologer. Skrifter 15

Massedigitalisering og edisjonsfilologi

Bidrag til en konferanse arrangert av
Nordisk nettverk for edisjonsfilologer

Voksenåsen, Oslo
2. til 4. september 2022

Utgitt ved
Annika Rockenberger
Aasta Marie Bjorvand Bjørkøy
Nina Marie Evensen
og Ellen Nessheim Wiger

Universitetsbiblioteket i Oslo
Oslo 2025

Innhold

Forord **2**
Massedigitalisering og edisjonsfilologi (*Annika Rockenberger, Aasta Marie Bjorvand Bjørkøy, Nina Marie Evensen, Ellen Nessheim Wiger*) . . . 3
Bidragene . . . 4
Presentasjon av konferansebidrag som ikke er blitt omgjort til artikkel 5
Takk . . . 10

Essay **13**
Nya vägar till det förflutna:
Två editionsfilologiska frågor till den storskaliga digitaliseringen (*Mats Malm, Ljubica Miočević*) . . . 14

Artikler **21**
Bibliography, database, and beyond:
Towards queer scholarly editing (*Jenny Bergenmar*) . . . 22
«Op ad Midten, krydser og falder af»
Dansemanuskriptet *XII Engels Danse* som kildeutgivelse (*Elizabeth Svarstad*) . . . 39
Digitale udgaver som HITL-garanter.
Om editionsfaglig domæneekspertberigelse og datakvalitetens betydning for integrationen af computationelle tilgange i humanvidenskaben (*Katrine Frøkjær Baunvig, Krista Stinne Greve Rasmussen*) . . . 50

Biografier **67**
Redaktører . . . 68
Forfattere . . . 69

Personregister **72**

Digitale udgaver som HITL-garanter. Om editionsfaglig domæneekspertberigelse og datakvalitetens betydning for integrationen af computationelle tilgange i humanvidenskaben

Katrine Frøkjær Baunvig, Krista Stinne Greve Rasmussen[1]
[1]Center for Grundtvigforskning | Aarhus universitet

Indledning

Less is more. I hvert fald hvis den mindre delmængde er af høj kvalitet. Sådan burde en tommelfingerregel lyde i de grene af humanvidenskaben, som gør beregningsorienteret brug af digitale data. Denne opfattelse er det vores hovedærinde at underbygge i denne artikel, som angår vigtigheden af den editionsfaglige berigelse af de kulturarvsdata, som ligger til grund for den mængde af digitale, videnskabelige udgaver (Digital Scholarly Editions, DSE), der i de seneste årtier har vokset sig omfangsrige i Norden.[1] Følgelig ønsker vi at pode det programmatiske akronym HITL (Human-In-The-Loop), som i de senere år i påtrængende grad har vokset sig relevant i fagfelter, der benytter sig af maskinlæring, på den editionsfaglige selvforståelse. 'HITL' anvendes om allehånde initiativer, hvor mennesker inddrages til at kvalitetssikre eller -forbedre data, der skal ligge til grund for en given algoritmisk model.[1] Det kan være til annoterings- eller kategoriseringsopgaver, som netop hører til de editionsfilologiske kerneopgaver: genreannotering, stedopmærkning eller personnavnsregistrering etc. I takt med at den Computationelle Humanioras virkefelt udvides, bliver pålidelige højkvalitetsdata stadig mere efterspurgte. Dette skærper relevansen af DSE-projekter og af de editionsfilologiske kompetencer, som sådanne projekter er baseret på.[2]

I det følgende vil vi med andre ord argumentere for, hvorledes klassiske editionsfilologiske fagdyder vinder ny relevans i en tid, hvor generativ AI og algoritmiske metoder i det hele taget sætter dagsordenen. I denne kontekst bliver humanvidenskabelig akkuratesse og kontekstforståelse attraktiv i alle vidensområder baseret på sprogteknologi og store sprogmodeller (Large Language Models, LLM). Det gør de, fordi sprogmodellers anvendelighed og pålidelighed er en direkte funktion af kvaliteten af de data, som de er trænet på. Som leverandører af kulturarvsdata af høj kvalitet forudser vi, at den editionsfilologiske fagstand vil indtræde som hjørnestenen i fremtidens AI-baserede vidensøkologi. I denne artikel vil vi imidlertid begrænse os til en illustration fra forskningsverdenen: vi vil give eksempler på, hvorledes den digitale editionsfaglige berigelse af ét

[1]Patrick Sahle redigerer et katalog over DSE's, hvorfra man fremsøge nordiske udgaver: `https://www.digitale-edition.de/exist/apps/editions-browser/$app/index.html`. Kataloget viser 18 nordiske udgaver (fordelt på 6 danske, 5 finske, 3 norske, 3 svenske og 1 islandsk), men kataloget er ufuldstændigt. Eksempelvis mangler de danske udgaver *H.C. Andersens Eventyr og Historier. Den digitale manuskriptudgave* (2020 f.), `https://hcams.andersen.sdu.dk/exist/apps/andersen/index.html`, og *Diplomatarium Danicum* (afsluttet), `https://diplomatarium.dk/`.

væsentligt dansk kulturarvsdatasæt *Grundtvigs Værker* har muliggjort computationel udforskning, som for sin del har kastet afgørende, nye forskningsresultater af sig. Med afsæt i en skitsering af 1) filologiens fundamentale position i humanioras udviklingshistorie med særinteresse for 2) editionsfilologiens placering i den humanvidenskabelige fagfamilie og forskningshistorie vil vi derfor i det følgende 3) beskrive de digitale editionsfaglige projekters nøglerolle i produktionen af højkvalitetskulturarvsdata for 4) at illustrere, hvorledes dette arbejde konkret baner vej for kvalitativt nye former for forskningsresultater.

2. Filologien er hjørnestenen i humanvidenskaberne og den Computationelle Humaniora

Få vil modsige, at filologien er udgangspunkt og centrum for humanvidenskaberne, som vi kender dem i dag. Filologien som begreb har været med til at forme udforskningen af mennesket med fokus på sprog og tekst siden Platon.[3]

Videnskabelig udgivelse er en helt særlig form for filologi, der har eksisteret stort set lige så længe. Formålet har altid været at gøre tekster tilgængelige på tværs af tid og sikre deres videre eksistens. Metoderne og de videnskabelige principper har varieret, ligesom man nu kan tale om forskellige filologiske grene afhængig af materialet, der gøres tilgængeligt. Den videnskabelige udgivelsesvirksomhed har i nyere tid desuden forandret sig fra at have været betragtet som en slags hjælpedisciplin for de øvrige videnskaber til at udgøre en selvstændig videnskabelig disciplin med et betydeligt teoretisk opsving i slutningen af det 20. århundrede.[4, s. 33] Videnskabelige udgaver har været en central del af nationsopbygningen, og de har inden for alle grene af humaniora tilvejebragt og formidlet vores tekstbårne historie.

Fokus i denne artikel er editionsfilologien, en gren af udgivelsesfilologien, der hovedsageligt beskæftiger sig med nysprogligt materiale fra efter bogtrykkets indførelse. Johnny Kondrup inddeler feltet i fem genstandsområder: «tekstkritik, variantapparater, dokumentbeskrivelse (bibliografi og manuskriptbeskrivelse), kommentering og elektronisk udgivelse».[5, s. 17] Inddelingen kan diskuteres, hvilket Kondrup også gør opmærksom på, men desuagtet er det passende, at tekstkritik nævnes som det første punkt, da det står som det helt centrale element i videnskabelig udgivelse, jf. også ovenfor. Overordnet set beskrives tekstkritikken «lämpligen som en samling verksamheter som utvecklats sedan antiken för att på metodisk grund välja bättre läsarter framför andra, när det föreligger textuell variation och/eller misstanke om korruption i det transponerade textmaterialet.»[6, s. 129-130] Bedre læsemåder skal ikke forstås æstetisk, men snarere hermeneutisk. Når tekster vandrer fra en medialisering til en anden, altså transponeres, kan der opstå fejl. Det kan både være egentlige fejl i teksten, f.eks. ord der ændrer form, men det kan også være materiel korrumpering. Afhængig af videnskabelig metode, tekstopfattelse og retningslinjer for transponering kan betydningen af disse korrumperinger vægtes forskelligt. Eksempelvis kan en transponering lægge stor vægt på den grafiske form og derfor søge at gengive de grafiske elementer så eksakt som muligt.

Centralt er opmærksomheden på, at tekster aldrig bare er tekster – og ikke mindst, at de aldrig kan gengives uden informationstab. Derudover kan tekster

også i sig selv indeholde fejl. «Det er et sundt filologisk (og hermeneutisk) princip at gå ud fra, at enhver konkret foreliggende tekst – en realtekst – indeholder fejl.»[5, s. 139] Realtekster er i denne terminologi teksten, som den fremtræder for os, hvorimod idealteksten er en idealisering. Emendationer uden et foreliggende tekstvidne med en variant gøres med udgangspunkt i forestillingen om en idealtekst. En længere teoretisk diskussion af tekstbegrebet ligger uden for denne artikels ramme. Derimod er det vigtigt at understrege, at videnskabelig tekstgengivelse har indbyggede teoretiske og metodiske refleksioner, der både er historisk foranderlige og – hvad vigtigere er – videnskabeligt forankrede, men altid afhængige af det menneske, der udgiver teksten.[5, s. 140][7, s. 20] Ingen computer kan forventeligt foretage komplekse emendationer, som når vi i Grundtvigs Værker har rettet «Almagt» til «Afmagt» i Grundtvigs artikel «Om Religion og Liturgie».[2] Derimod kan man forvente, at man med AI kan træne en model til at rette mere trivielle tekstrettelser, for eksempel omvendte bogstaver. Mennesker må altså komplementere maskiner for retteligt at udøve den tekstkritik, der ligger til grundt for kvalificeret reproduktion og -distribution af tekster.

En central af det videnskabelige udgivelsesarbejde, som på sin vis falder inden for tekstkritikken, men går forud for denne, er selve tilvejebringelsen af tekster ved hjælp af scanning og tekstgenkendelse. I Grundtvigs Værker kollationerer vi den digitale version af grundteksten tre gange, både for at identificere varians, men i særdeleshed også for at genoprette den automatiske tekstgenkendelse, så vores digitale tekster er i overensstemmelse med det trykte forlæg. Ingen automatisk tekstgenkendelse er endnu fejlfri, og selvom det kan virke som et triviel skridt at kollationere, bliver denne del af tekstkritikken først rigtig kvalificeret i kraft af editionsfilologen, der også kan identificere eventuelle tekstfejl uden forlæg. Nøjagtigheden af genkendelse er dog ikke kun relevant for den kvalificerede tekstgengivelse, den er også afgørende for den potentielle brug af teksterne: «The accuracy of Optical Character Recognition (OCR) technologies considerably impacts the way digital documents are indexed, consulted and exploited.»[10] Det har været almindeligt at opfatte en vis mængde støj som irrelevant for udnyttelsen af tekster inden for den Computationelle Humaniora, eksempelvis til fjernlæsning ('distant reading'), men ved at acceptere støj risikerer man at lukke af for visse typer forskningsspørgsmål og -interesser. Selvom et korpora fuld af støj statistisk set kan være nok for indeværende, er det ikke nødvendigvis også sandt i fremtiden. En bedre OCR-proces kan ikke blot have betydning for den tekstnære forskning, men også påvirke læsere til en mere positiv, og dermed brugbar, opfattelse af materialet.[11] Således handler opretning af OCR-fejl ikke kun om materialet selv, men om hele den videnskonomi og -økologi teksterne indgår i.

Af editionsfilologiens fem genstandsområder, som blev omtalt tidligere, skal yderligere to behandles her, nemlig dokumentbeskrivelse og kommentering. Førstnævnte skal ideelt set gøre teksten nærværende for brugeren, men de færreste har nok den oplevelse, når de eksempelvis læser en bibliografisk formel eller en manuskriptbeskrivelse. Derudover vil mange, der lægger det fysiske dokumenter til grund for studier, behøve at se og arbejde med dokumentet selv.

[2]Eksemplet blev først brugt i en forelæsning af Johnny Kondrup på Københavns Universitet. [8, s. 146] Se også [9]

Dokumentbeskrivelsen kan med andre ord ikke erstatte det fysiske eksemplar. Alligevel indsamler dokumentbeskrivelsen værdifuld metadata, der kan bruges i større digitale udforskninger at materialesamlinger. For *Grundtvigs Værkers* vedkommende kunne dette være oplysninger om tidsskrifters størrelser (fysisk og oplag) og redaktører. Dokumentbeskrivelser er derfor ikke blot beskrivende, men også berigende i forhold til det digitale korpus en videnskabelig udgave også er.

Endnu mere berigende er kommenteringen, som vi vil demonstrere med eksempler fra forskningsstudier nedenfor. I *Grundtvigs Værker* har vi to typer kommentarer: overgribende kommentarer og punktkommentarer. Den overgribende er indledningerne og tekstredegørelserne, mens punktkommentarerne består i verbalkommentarer (der udarbejdes lokalt for hvert værk) og registeropmærkninger (der er globale i udgaven, dvs. identisk gentaget overalt for samme lemma). I registeropmærkningerne findes når personer, steder, titler, mytologiske entiteter og bibelreferencer nævnes i teksterne. Med et specifikt id-nummer henvises til en central fil med globale kommentarer, der vises lokalt i tekststedet. Dette gør redigering af posterne lettere og sparer tid ved størrelser, der allerede findes forklaringer på. Kommentering er, ved siden af tekstkritikken, editionsfilologiens kerneopgave, der i den videnskabelige udgave har til opgave af bygge bro over den historiske kløft mellem tekster og læsere, men som også i forskningen kan udnyttes til computationelle studier. Vi vil nu se nærmere på editionsfilologiens rolle mere bredt i etableringen og udforskningen af digital kulturarvsdata.

3. DSE som HITL i produktionen af højkvalitetskulturarvsdata

2002 anslås ofte som begyndelsespunktet for den såkaldt digitale æra. I årene umiddelbart op til og umiddelbart efter 2010 udmøntede det sig i begrebet *Big Data*. Herom vidner søgninger på Google Trends, i forskningslitteraturdatabaser og på platforme, der tilgængeliggør offentlige mediedata. Begrebet voksede sig til i datalogiske fagmiljøer og dækker over datasæt, hvis størrelse og kompleksitet omkalfatrede traditionelle metoder til indsamling, opbevaring, behandling og fortolkning af data. Et omsiggribende internet førte til en betydelig forandring af målestokken for tilgængelige data; i disse år gik man fra at tælle data i gigabytes (109) til at opgøre dem i petabytes (1015). Udviklingen i størrelse ledsagedes af en udvidelse af datamaterialets variation samt en drastisk forøgelse af behandlingshastigheden. Det oplevedes som et væsentligt faseskifte.

Begejstring og videnskabelig optimisme omgærdede denne oplevelse og begrebet spredte sig nu på vingerne af ånden i den irske forfatter Oscar Wildes udbredte *bon mot*: «Nothing succeeds like excess». Skønt begrebet Big Data ikke har nogen formel definition voksede det med andre ord frem som et løfte. De nye rammer for datamængder, datavariation og databehandlingshastighed forventede man ville vise sig som løsningen på en broget mængde opgaver – af såvel videnskabelig som industriel, samfundsmæssig og kulturel natur – man ikke før var i stand til at tage sig af. Selvom man stadig kan støde på annoncer, der holder fast i den begejstrede tone og eksempelvis ønsker at redegøre for «How Big Data Can Help You Do Wonders In Your Business»,[12] er den generelle entusiasme aftaget og begrebet afmattet. Det skyldes først og fremmest, at forventningerne til de afgørende nybrud er blevet

omdirigeret til kunstig intelligensteknologien – ikke mindst til den generative AI.

Denne drejning er også et resultat af, at de massive datamængder, som Big Data-hypen tog afsæt i, var kendetegnet ved en anseelig grad af rod eller skidt. Som tiden gik, stod det altså klart, at Big Data i et vist omfang var synonymt med *Dirty Data*. Det vil sige, at mega-repositorierne som hovedregel indeholdt ufuldstændige, upræcise eller ligefrem ukorrekte data. I første omgang var den generelle holdning, at dataomfanget i sig selv udlignede skidtet eller at en automatiseret rensning kunne sikre dataintegriteten. Men gradvist stod det klart, at dette ikke var tilfældet. Og i en lang række faglige og industrielle sammenhænge er det efterhånden blevet uomtvisteligt, at pålidelige og bæredygtige data kræver menneskeligt tilsyn og håndtering.

Denne indsigt har fremelsket akronymet, der gemmer sig i denne artikels overskrift: HITL eller Human-In-The-Loop. Det anvendes i en righoldig mængde kontekster, hvor automatisering har vist sig at skabe praktiske såvel som etiske udfordringer. HITL har tilstrækkelig spændvidde til også at omfatte det editionsfaglige bidrag til produktionen og vedligeholdelsen af højkvalitetskulturarvsdata. Editionsfilologer indtager i kraft af deres møjsommelige arbejde med at tilvejebringe og kuratere digitale udgaver rollen som humans-in-the-loop – som den menneskelige valideringsinstans i den digitale kulturarvsbevarings kredsløb. I den forstand bør de opfattes som det første led i den kæde, som er den fremvoksende computationelle humaniora består af. Uden menneskelig fejlsikring, opmærkning eller anden berigelse af de relevante datasæt, er det ikke rimeligt at stille disse datasæt hermeneutisk komplekse spørgsmål. Med andre ord: uden denne sikring af en given datasamlings kvalitet vil udsagnskraften af en beregningsbaseret tekstanalyse, som bygger herpå, være begrænset.

Uden at fortabe os i spørgsmålet er det på sin plads at fremdrage en oplagt og uheldig konsekvens af disse omstændigheder: Menneskelig validering af et datasæt er en omkostningsrig affære – DSE-projekter kræver mange ressourcer i form af tid og penge. Offentlige og private fonde vil derfor naturligt nok være særdeles kritiske med, hvilke datasæt, de ønsker at understøtte berigelsen af. I praksis betyder det, at det gerne er i forvejen kendt og anerkendt materiale, som modtager støtte. Med andre ord flyder der en konservativ kanon- eller kulturreificerende tendens ud af kravet om HITL og højkvalitetsdata. For at skære situationen hård op: det er oftere projekter, som tager afsæt i kulturbærende figurer som N.F.S. Grundtvig, Zacharias Topelius, Henrik Ibsen, August Strindberg, der modtager støtte end projekter, der kredser om ukendte arbejderkvinders efterladte papirer. Logikken, der driver dette forhold frem, er forståelig; vi håber ikke desto mindre, at fremtidige DSE-ansøgere såvel som DSE-bevillingsgivere vil forsøge at modvirke de åbenlyse misforhold i kulturarvsbevaringen, som følger af tankerefleksen og bevillingsimpulsen. Derfor har vi på Center for Grundtvigforskning søgt at udvide vores dataetableringspligt fra Grundtvigs forfatterskab til hele den brogede og mindre prominente brugskultur, som forfatterskabet satte i gang. Ét håndgribeligt resultat af disse bestræbelser er et af *Grundtvigs Værkers* ledsageprojekter: en højkvalitetsdigitalisering af det grundtvigsk sindede tidsskrift *Højskolebladet* (1876-2020), som vi i 2021 modtog midler fra Carlsberg Fondet til at digitalisere. I nær fremtid vil korpusset være at finde på Det Kongelige Biblioteks side tidsskrift.dk.

Begge samlinger er højkvalitetsmateriale, der muliggør udforskning af Grundtvigs liv, værk og efterliv på måder, som hidtil ikke har kunnet lade sig gøre. Det er eksempler på denne udforskning og dens brug og afhængighed af en DSE, som vi nu skal give et indblik i. Vi har begrænset os til to eksempler. Det første angår et tvillingestudie af Grundtvigs emfasetilskrivninger ved hjælp af henholdsvis spatiering og udråbstegn; det andet angår en begrebsmæssig efterprøvning af Grundtvigs farvebrug.

4. Betydningen af DSE for den computationelle humanvidenskab: Eller om udråbstegn og farvebegreber i *Grundtvigs Værker*

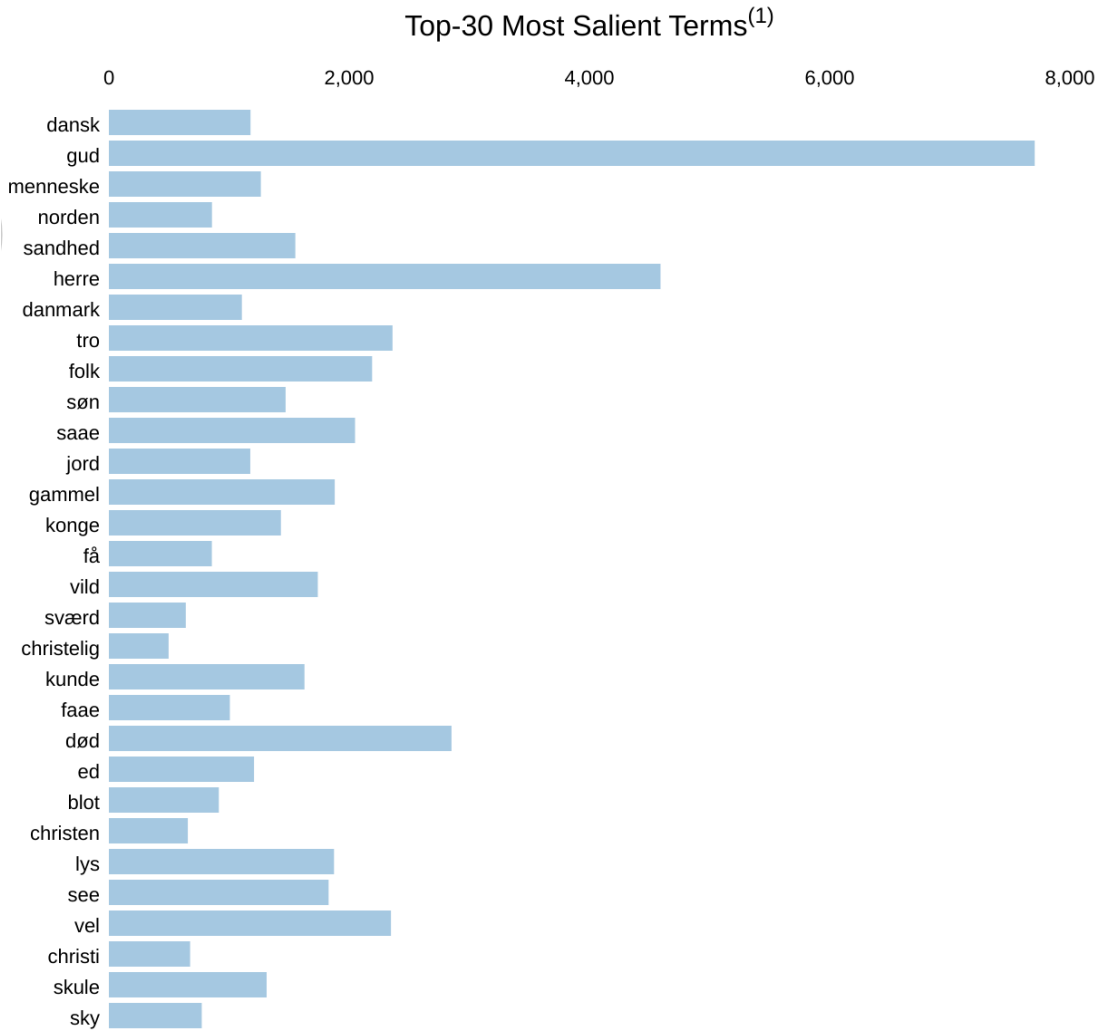
Det første demonstrationseksempel vi har valgt er metodisk simpelt og bygger først og fremmest på to specifikke træk ved *Grundtvigs Værkers* opmærkning (emfase-studiet). Det andet eksempel baserer sig på en matematisk mere kompleks fremgangsmåde – såkaldt simple neurale netværk eller *word embeddings* – der for sin del inddrager alle niveauer af opmærkningsarbejdet (farve-studiet). Der er med andre ord tale om to særdeles DSE-afhængige computationelle studier omend afhængigheden gestalter sig på meget forskelligartet vis. Begge studier har bragt Grundtvigforskningen til klarhed over hver sit felt, hver sit spørgsmål. Emfase-studiet har blotlagt, hvad der var de semantiske tyngdepunkter i forfatterskabet, mens farve-studiet har efterprøvet, om Grundtvig selv efterlevede sin berømte regel: «Med Farver kan man ingen Christus male».[13, s. 171] Begge undersøgelser har forlenet forskningen med ikke-trivielle indsigter og beriget vores forståelse af Grundtvigs tankeverden.

4.1 De kristne eftertryk og de danske betoninger

De to artikler «Emotional Imprints: Letter-Spacing in N.F.S. Grundtvig's Writings» og «Emotional Imprints: Exclamation Marks in N.F.S. Grundtvig's Writings»[14][15] tager afsæt i den indvending, man undertiden støder på i Grundtvig-forskningen og -formidlingen, at Grundtvigs kernesager ikke flugter med receptionen af ham som kirkefader og landsfader[16, s. 92-99] – at eftertiden ikke har begrebet ham og hans ærinde. Tvillingeundersøgelserne tillader sig at tage denne påstand på ordet ved at teste den. De tager desuden begge afsæt i to forhold: 1) det digitale, opmærkede Grundtvig-materiale og 2) den præmis, at Grundtvig var påvirket af en romantisk 'emotionaliseringsstrømning', der var kendetegnet ved en bestræbelse på at mætte de fleste udtryksformer – også tryksager – med følelser. Emfasetilskrivning som s p a t i e r i n g, der som fremhævelsesmiddel var frakturens parallel til antikvaens kursivering, og udråbstegn var de mest almindelige redskaber hertil – redskaber, som Grundtvig var en endog særdeles flittig bruger af.

Spatieringerne er konsistent opmærket i hele Grundtvigs Værker, hvilket reducerede den tekniske del af undersøgelsen til et simpelt udtræk. Vi undersøgte distributionen over tid og muligheden for påvirkning fra skiftende forlæggere. Begge dele syntes dog ikke at vise nogen påfaldende endsige signifikante tendenser. Fordelingen viser med andre ord ikke noget klart mønster. Det samme gør sig gældende for Grundtvigs brug af udråbstegn, som også meget simpelt lod sig

udtrække. Hvad der imidlertid tegner et klart mønster for både spatieringer og udråbstegn, er de termer, som emfaseredskaberne knytter sig til. Grundtvig understreger og udråber på den ene side ord, der knytter sig tæt til Danmark og til kristendommen – til 'dansk' samt til 'Jesus' og 'Kristus'. Denne påstand underbygges af nedenstående tabeller.



Figur 1. Grundtvigs 30 mest brugte ord i sætninger, der munder ud i udråbstegn.

I figur 1 ser vi de ord, som hyppigst indgår i sætninger, der munder ud i et udråbstegn. I figur 2 ser vi listen over de hyppigst spatierede termer. Det fremgår heraf, at der er to tyngdefelter i Grundtvigs tænkning. At Grundtvig sidenhen forbindes med tanker om danskhed og om dansk kristendom må på den baggrund siges at være i overensstemmelse med det indlejrede budskab i værkerne.

4.2 Grundtvigs farvebrug

Grundtvig var billedskeptisk, men han var samtidig overmåde farveglad. Det er udgangspunktet for en undersøgelse, der er grundlaget for artiklen «'Med Farver kan man ingen Christus male'. Joakim Skovgaard og Grundtvigs farver».[17]

Artiklen forsøger at besvare, om den farveglade Grundtvig ligefrem holder Kristus fri af farverne i sine sprogbilleder, når han nu er så forhippet på at oplyse sine kunstnerkolleger om, at billedkunst og kristendom ikke har noget med hinanden at skaffe. Til forskel fra de to forrige undersøgelser gør denne undersøgelse brug af stor regnekraft og er baseret på en algoritme trænet på det samlede digitaliserede korpus bestående af de over 8 millioner ord fordelt på over 1000 værker, som Grundtvig udgav i løbet af en periode på 68 år (1804-1872). På samme måde, som et landkort komprimerer og forenkler information fra et virkeligt landskab,[18] kan Grundtvig-algoritmen forstås som en sammenpresning og forenkling af det komplekse og multidimensionale rum, Grundtvigs forfatterskab og ordforråd repræsenterer. Både kort og algoritme kan betragtes som navigationsudstyr: begge muliggør orientering i et flerdimensionalt rum. I denne sammenhæng sætter algoritmen forskere i stand til at udarbejde såkaldte simple neurale netværk kaldet word embeddings, som svarer til kortlægningen af et ords associationsstruktur i et bestemt tekstkorpus.[19][20] Denne fremgangsmåde hjælper i denne sammenhæng til et overblik over de farver, hvormed Grundtvig i sine sprogbilleder fremmaler sin forståelse af verden, af naturen, af folkefærd. Den editionsfaglige møjsommelige oprensning og opmærkning af Grundtvig-materialet er nødvendig, så der ikke opstår fejl eller støj i kompressionen eller disproportioner på associationskortet. Undersøgelsen af Grundtvigs farvebegreber har hjulpet til nye indsigter i forskningen.

Denne visualisering af en *word embedding* viser, hvorledes 'Christus' forholder sig til serien af *blaa*, *grøn*, *gul*, *rød*, *brun*, *sort*, *hvid*, *guld* og *sølv*. Her ser vi tydeligt, at 'Christus' ikke danner betydningsforbindelser med de jordiske farver. I stedet lægger han sig tilrette på i en akavet treenighed, hvor han har fået følgeskab af med sin egen jordiske modpart Jesus samt Vorherre. Denne treenighed er 'akavet', fordi Helligånden mangler. Den står ellers centralt i Grundtvigs tænkning.[21] At den er helt fraværende fra hans farveassociationer er derfor bemærkelsesværdigt. Visualiseringen hjælper til at vise, at Kristus og særligt Helligånden er uden for farvernes rækkevidde. Det er de antageligvis, fordi de begge to først og fremmest er lydfænomener – Grundtvig fremstiller ligefrem ånden som en klangkuppel, hvor tonerne fra den himmelske oven og fra fortiden runger og resonnerer.[21] Når Grundtvig skal beskrive den guddommelige treenighed, styres han med andre ord af en akustisk logik – der fx er knyttet til Johannesevangeliet 1,14 og forståelsen af Jesus som «Ordet», der «blev kød». Og det er givetvis derfor han oftere slår ud i de larmende sprogbilleder fremfor de farverige, når han skal beskrive dem. Denne indsigt, som er faciliteret af Grundtvig-algoritmen, er ikke før blevet ført så håndfast igennem.

5. Afslutning

Vi er overbeviste om, at editionsfilologien har en afgørende og central rolle at spille i humaniora generelt og den compuationelle humaniora specifikt. Editionsfilologiske kompetencer er værdifulde i den digitale transformation af tekster, der har stået på længe, men som nu ikke længere er en særeksistens for tekster. Nu *er* alle tekster digitale.

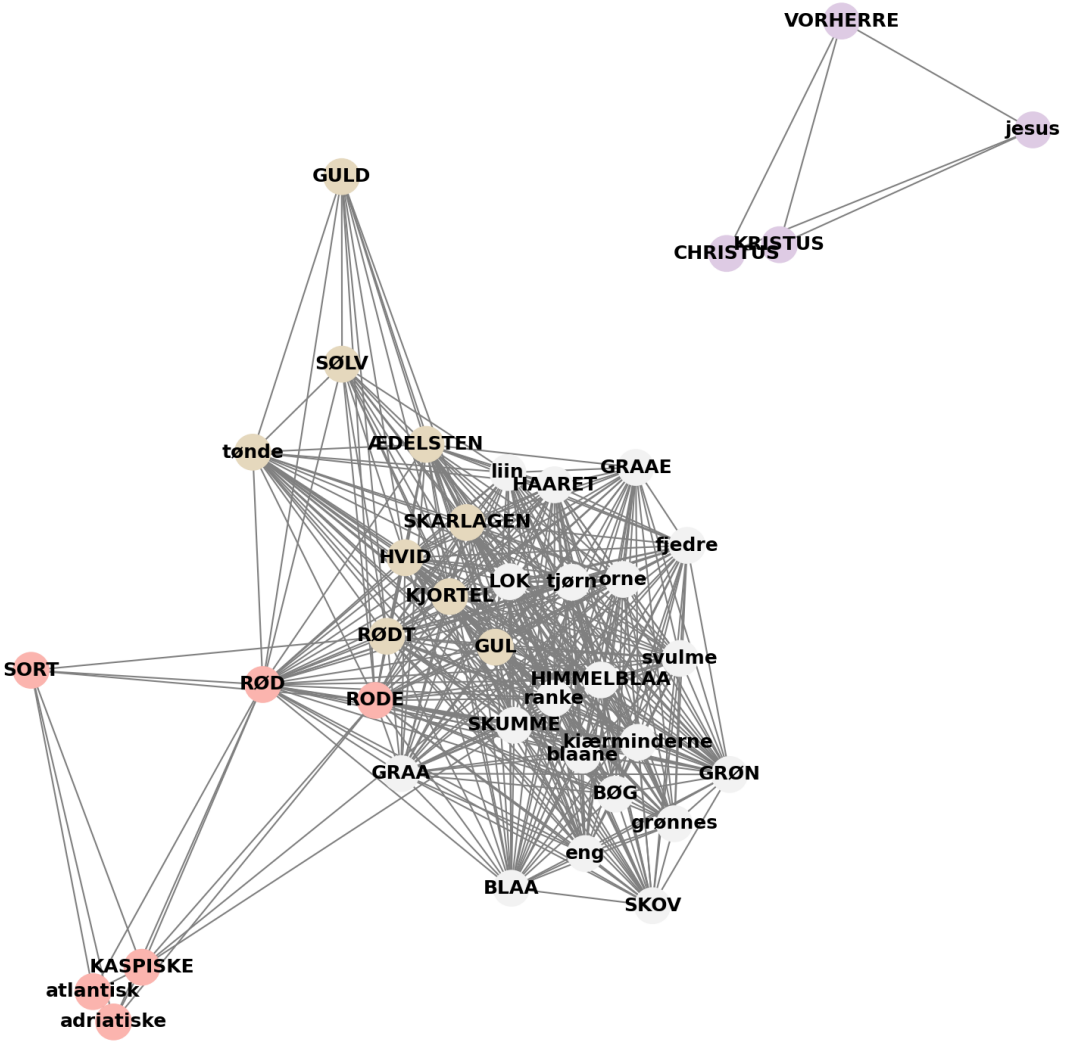
Studierne vi henviser til i denne artikel, er mulige at foretage på tekst, der blot er

OCR-behandlet, men pålideligheden af undersøgelsernes resultater afhænger af, at teksterne er blevet rensede, dvs. har været igennem menneskelige hænder. Ligeledes kan de her nævnte studiers brug af *word embeddings* i princippet gennemføres uden de videnskabelige udgavers opmærkede tekster, men dette finder vi ikke attraktivt. Det er vigtigt, at materialet har høj standard. Eksempelvis blev emfasestudiet gennemført på et tidspunkt, hvor blot 40 % af GV-korpusset var renset og opmærket, hvilket giver fejl som lemmaet «få» i figur 1, der rettelig burde være «faa». Jo længere vi er med udgaven, jo bedre bliver korpus, og jo bedre fungerer modellerne. Principielt kunne man til disse formål argumentere for, at et renset OCR-korpus var rigeligt, men så ville studierne ikke kunne gå videre end dette. Aktuelt arbejder vi med at træne taggers, der blandt andet muliggør nye former for automatiseret opmærkning, men – og dette er en vigtig pointe – nye studier, hvis metoder vi endnu ikke kender. Muligheder vi opfatter som betinget af editionsfilologisk ekspertise og arbejde.

Table 2. 20 Most Frequent Letter-Spaced Words.

	L-S Words	L-S Words (English)	Total Amount of Occurrences	First Occurence in Text
1	Danmark	Denmark	1106	78
2	Danske	Danish	1105	51
3	ikke	not	735	16
4	Jesus	Jesus	601	62
5	Daaben	The baptism	463	78
6	dansk	Danish	451	47
7	Jesus Christus	Jesus Christ	449	82
8	Damerne	The ladies	444	5
9	Ordet	The word	434	80
10	Odin	Odin	422	13
11	levende	living	415	49
12	Aanden	The spirit	415	57
13	det	it	414	15
14	Norden	The North	401	27
15	den	it	383	18
16	Danmarks	Denmark's	380	43
17	Frihed	Freedom	380	69
18	Thor	Thor	372	21
19	Zeus	Zeus	337	5
20	Jesu	Jesus'	328	28

Figur 2. Grundtvigs 20 hyppigst spatierede ord.



Figur 3. Det semantiske væv, der binder termerne 'Christus' og blaa, grøn, gul, rød, brun, sort, hvid, guld samt sølv sammen i Grundtvigs publicerede forfatterskab.

Bibliografi

[1] Vikram Singh Bisen. *What Is Human in the Loop Machine Learning: Why & How Used in AI?* VSINGHBISEN, May 20, 2020. URL: https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0..

[2] Katrine Frøkjær Baunvig. "A Computational Future? Distant Reading in the Historical Study of Religion". In: *Stepping Back and Looking Ahead: Twelve Years of Studying Religious Contact at the Käte Hamburger Kolleg Bochum*. Ed. by Maren Freudenberg et al. Leiden: Brill, 2023. DOI: 10.1163/9789004549319_013.

[3] James Turner. *Philology: The Forgotten Origins of the Modern Humanities*. Princeton: Princeton University Press, 2014. DOI: 10.1515/9781400850150..

[4] Johnny Kondrup. "Indledning til Dansk Editionshistorie". In: *Dansk Editionshistorie. Udgivelse af græsk og latinsk litteratur*. Ed. by Christian Troelsgård and David Bloch. Vol. 1. København: Museum Tusculanums Forlag, 2021, pp. 31–48.

[5] Johnny Kondrup. *Editionsfilologi*. København: Museum Tusculanums Forlag, 2011.

[6] Mats Dahlström. *Under utgivning. Den vetenskapliga utgivningens bibliografiska funktion*. Borås: Valfrid, 2006.

[7] David Greetham. "A History of Textual Scholarship". In: *The Cambridge Companion to Textual Scholarship*. Ed. by Neil Fraistat and Julia Flanders. Cambridge: Cambridge University Press, 2013, pp. 16–41. DOI: 10.1017/CCO9781139044073.002..

[8] Nikolai Frederik Severin Grundtvig. "Om Religion og Liturgie". In: *Ny Minerva* (Feb. 1807), pp. 129–33. URL: http://www.xn--grundtvigsvrker-7lb.dk/tekstvisning/650/0#..

[9] Kirsten Vad and Krista Stinne Greve Rasmussen. "Kommentarer, registre og data". In: *Kommentering og præsentation af nordiske tekster i en digital tidsalder*. Ed. by Simon Skovgaard Boeck, Karen Skovgaard-Petersen, and Jeppe Barnwell. København: Det Danske Sprog- og Litteraturselskab, upcoming.

[10] Guillaume Chiron et al. "Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information". In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Joint Conference on Digital Libraries. 2017, pp. 1–4. DOI: 10.1109/JCDL.2017.7991582. URL: https://doi.org/10.1109/JCDL.2017.7991582.

[11] Kimmo Kettunen et al. "Optical Character Recognition Quality Affects Subjective User Perception of Historical Newspaper Clippings". In: *Journal of Documentation* 79.7 (2023), pp. 137–56. DOI: 10.1108/JD-01-2023-0002. URL: https://doi.org/10.1108/JD-01-2023-0002.

[12] Simplilearn. *How Big Data Can Help You Do Wonders In Your Business*. URL: https://www.simplilearn.com/how-big-data-can-help-do-wonders-in-business-rar398-article.

[13] Nikolai Frederik Severin Grundtvig. "Kiærminde-Bladet". In: *Danne-Virke et Tids-Skrift* 3 (Oct. 1817), pp. 155–72. URL: http://www.xn--grundtvigsvrker-7lb.dk/tekstvisning/18413/0#.

[14] Katrine F. Baunvig, Oliver Jarvis, and Kristoffer L. Nielbo. "Emotional Imprints: Letter-Spacing in N.F.S. Grundtvig's Writings". In: *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, CEUR*. Digital Humanities in the Nordic Countries 5th Conference. Vol. 2612. 2020, pp. 192–202. URL: https://ceur-ws.org/Vol-2612/short3.pdf.

[15] Katrine F. Baunvig, Oliver Jarvis, and Kristoffer L. Nielbo. "Emotional Imprints: Exclamation Marks in N.F.S. Gruntvig's Writings". In: *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries, CEUR*. Digital Humanities in the Nordic Countries. Vol. 2865. 2021, pp. 156–169. URL: https://ceur-ws.org/Vol-2865/short7.pdf.

[16] Jes Fabricius Møller. *Grundtvigs død*. Aarhus: Aarhus Universitetsforlag, 2019.

[17] Katrine Frøkjær Baunvig. "'Med Farver kan man ingen Christus male'. Joakim Skovgaard og Grundtvigs farver". In: *Kraft og skønhed. Joakim Skovgaards udsmykning i Viborg Domkirke*. Ed. by Iben Overgaard. Viborg/København: Skovgaard Museet: Strandberg Publishing, 2023, pp. 228–239.

[18] Alfred Korzybski. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. Lancaster: Penn, 1933.

[19] Vincent D. Blondel et al. "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008 (Oct. 2008). DOI: 10.1088/1742-5468/2008/10/P10008.

[20] Xin Rong. *Word2vec Parameter Learning Explained*. arXiv. Nov. 1, 2014. URL: https://doi.org/10.48550/arXiv.1411.2738.

[21] Katrine Frøkjær Baunvig. “Ånd hos Grundtvig: En digitalhumanistisk skitse af begrebsbrugen i Grundtvigs værker”. In: *Åndsfrihed: i det moderne samfund*. Ed. by Ove Korsgaard and Michael Agerbo Mørch. Aarhus: Aarhus Universitetsforlag, 2022, pp. 65–102.
