Winnie Collin, Inger Schoonderbeek Hansen, Tina Thode Hougaard og Karen Schriver (udg.):
20. Møde om Udforskningen af Dansk Sprog, Århus 2024

DanDIGI – udvikling af et korpus med dansk digitalt medieret interaktion

Philip Diderichsen & Torben Juel Jensen
Københavns Universitet

1. Indledning – hvorfor er der behov for et korpus?

Sociale medier, dvs. web- og mobilbaserede applikationer der gør det muligt for brugere at skabe, dele og cirkulere brugergenereret indhold (Lomborg 2025), har i perioden siden årtusindskiftet markant ændret måden vi bruger skriftsprog på i hverdagen. Studier af digitalt medieret interaktion (også kaldet *computer-mediated communication, CMC*; Jensen 2013) er derfor afgørende både for at forstå vores sociale liv i dagligdagen og for at generere og teste hypoteser om sprogforandring, herunder ikke mindst om hvordan sprogbrugen reagerer på mediets *affordances* (handlingsmuligheder; Gibson 1977; Hutchby 2001) med hensyn til blandt andet dialogicitet, opfattet formalitet og multimodale ressourcer.

Det er da heller ikke fordi det skorter på studier af interaktion og sprogbrug på sociale medier, heller ikke i en dansk kontekst. Det er imidlertid kendetegnende for dem at de falder i to grupper: 1) De er baseret på et relativt lille datamateriale som blandt andet pga. indsamlingsformatet (ofte billedfiler) mest egner sig til kvalitative undersøgelser, eller 2) de baserer sig på datamateriale som nok er omfangsrigt men pga. forskningsinteresser og/eller anonymiseringshensyn indsamlet uden at bevare information om hvordan interaktionen forløber (dvs. hvordan en post evt. indgår i en længere tråd), om deltagerne, tid og (under)forum. Samtidig er det af

indsamlingstekniske og juridiske årsager ofte meget ensidigt i forhold til hvilke sociale medier som dækkes.[1]

En afgørende grund til denne situation er efter vores vurdering at der i modsætning til hvad der er tilfældet for talt dansk og skriftligt dansk fra traditionelle medier – fx LANCHART-korpusset (Sprogforandringscentret u.å.) og KorpusDK (DSL u.å.) – i øjeblikket ikke eksisterer et bredt sammensat korpus af dansk sprogbrug på sociale medier som er alment tilgængeligt for forskere. Man kunne ellers umiddelbart tro at det ville være nemt at etablere sådan et korpus i og med at data fra sociale medier er født digitale i modsætning til talesprog, der først skal optages og transskriberes, og skriftsprog i papirbaserede medier, som ligeledes kræver omfattende behandling før det foreligger i et pålideligt digitalt format. I praksis er et korpus baseret på sociale medier dog langtfra ligetil at etablere, og det er formentlig også derfor der ikke eksisterer et sådant i forvejen, i modsætning til hvad der er tilfældet for en række andre europæiske sprog (Frey et al. 2020). Grundene til dette er for det første GDPR-reglerne (EU 2016) idet der i interaktion på sociale medier ofte indgår mange deltagere som i nogle tilfælde skal give tilsagn før deres bidrag kan indsamles. For det andet sker interaktionen på kommercielle medieplatforme som i stigende grad gør automatiseret indsamling umuligt eller meget dyrt.

Projektet *Danish Digitally Mediated Interaction* (DanDIGI) har til formål at etablere et bredt sammensat korpus som gør det muligt at benytte såvel kvantitative som kvalitative metoder i studier af sprogbrugen på sociale medier. Det vil sige korpuslingvistiske metoder som søgninger efter sproglige strukturer og opstilling af konkordanser samt statistisk baserede beskrivelser og sammenligninger af tekster/delkorpora, men også nærsproglige analyser af sproglige strukturer i deres interaktionelle og multimodale kontekst. Samtidig skal projektet gerne danne grundlag for at korpusset kan udvides med data fra andre projekter, ved at grundlægge en solid digital infrastruktur. Projektet huses af Sprogforandringscentret på Københavns Universitet (www.dgcss.hum.ku.dk) og har som deltagere Tanya Karoli Christensen,

[1] Dette gælder fx Gigaword-korpusset og det materiale (daTenTen) som ligger tilgængeligt via SketchEngine (Derczynski et al. 2021; Jakubíček et al. 2013).

Philip Diderichsen, Torben Juel Jensen, Andreas Candefors Stæhr og Liisa Deth Theilgaard.[2]

Artiklen beskriver i høj grad *work in progress* idet projektet løber i perioden 2024-2026. Vi vil i det følgende beskrive først de overordnede principper i opbygningen og struktureringen af korpusset samt det valgte datagrundlag. Herefter vil vi fokusere på problemstillinger i den konkrete behandling og strukturering af data med henblik på at gøre dem klar til integration i tekstkorpusset. Endelig vil vi skitsere de nye former for undersøgelser som korpusset vil muliggøre.

2. Overordnede principper for DanDIGI-korpusset

2.1. Udnyttelse af allerede indsamlede data

Grundideen i DanDIGI er at udnytte og tilgængeliggøre datamateriale som er indsamlet i forbindelse med andre projekter. Der er altså tale om et *opportunistisk* korpus (McEnery & Hardie 2012: 11), men ud fra det materiale vi har adgang til, udvælger vi, i det omfang vi ikke har mulighed for at tage alt med[3], sådan at korpusset bliver sammensat bredest muligt. Det gælder i forhold til typer af sociale medier og til interaktionens placering på en skala gående fra kommunikation i helt åbne fora mellem individer der i vidt omfang ikke kender hinanden, over lukkede grupper til helt privat kommunikation mellem enkeltindivider.

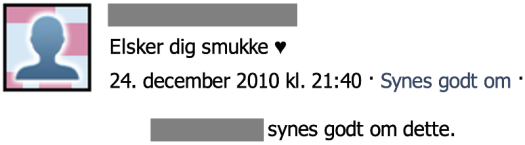
2.2. Ens struktur på tværs af platforme og projekter

Det er som nævnt et vigtigt formål med korpusset at kunne sammenligne sprogbrugen på tværs af datatyper. Derfor er det afgørende at korpusset bliver struktureret sådan at man kan søge og optælle i det på tværs af de forskellige platforme data er indsamlet fra, og på tværs af de projekter der oprindeligt har indsamlet dataene. En særlig problemstilling i den forbindelse er at data fra sociale medier er meget komplekse, dels ved at være udpræget multimodale (de indeholder billeder, emojis, links mv.), dels ved at bestå af en blanding

[2] Projektet er støttet af Carlsbergfondet (forskningsinfrastrukturmidler, grant CF23-1008).

[3] En del af det tilgængelige datamateriale kræver, som det vil blive beskrevet i afsnittet "Datagrundlag", manuel behandling hvorfor det ikke er muligt inden for projektets rammer at klargøre det hele til korpusset.

af brugergenererede bidrag og tekst genereret af platformene (dvs. diverse tidsangivelser, brugernavne, like-optællinger og lignende). Det er på den ene side afgørende at få adskilt den systemgenererede tekst fra den brugergenererede tekst da det jo er sidstnævnte vi er interesserede i at kunne undersøge. På den anden side udgør den systemgenererede tekst i mange tilfælde vigtige metadata i forhold til at kunne forstå og kontekstualisere interaktionen. Den kan derfor ikke bare slettes, men må markeres (opmærkes) sådan at den kan skelnes fra den brugergenererede tekst, og sådan at forskellige typer af information kan genkendes på tværs af medier og projekter. Eksempelvis er der både i opslag på Facebook og på Heste-Nettet angivet brugernavn, tid og "synes godt (hhv. dårligt) om"-optællinger foruden selve beskeden fra brugeren, men denne information (som fx vil være relevant i forhold til at sammenligne samme sprogbrugers sprogbrug over tid og evt. i forskellige medier, og hvilke typer af posts som får flest reaktioner) har hverken samme sproglige form eller helt samme placering i opslagene; se skærmklippene i figur 1 og 2 nedenfor. Vi udfolder disse eksempler mere nedenfor.



Figur 1. Facebookopslag fra 2010.

Figur 2. Opslag på Heste-Nettet.

2.3. FAIR-principperne for forskningsdatamanagement

Da DanDIGI-korpusset er tænkt som infrastruktur for fremtidig forskning, dvs. en ressource som gerne skal kunne bruges af mange personer, i forskellige øjemed og i lang tid fremover, følger vi FAIR-principperne for forskningsdatamanagement (Wilkinson et al. 2016). FAIR står for *Findability*, *Accesibility*, *Interoperability* og *Reusability*.

Data skal altså for det første være mulige at *finde* og være *tilgængelige* for andre forskere. Dette sikrer vi primært gennem registrering og (i et vist omfang) deponering af korpusset i CLARIN-infrastrukturen[4] samt tilgængeliggørelse via Sprogforandringscentrets korpusinfrastruktur.

For det andet skal både data og metadata foreligge i et formaliseret, veldefineret og velbeskrevet format som gør at de er mulige at anvende i andre sammenhænge (*reusability*) og i andre it-systemer (*interoperability*) end vores. Her er der flere muligheder som opfylder det grundlæggende krav om at være veldefineret og maskinlæsbart, hvilket er helt nødvendigt for automatisk at kunne processere de store datamængder og overføre dem til vores korpusværktøj og andre ønskede formater: tabelbaserede dataformater (fx kommasepareret tekst (CSV) eller relationelle databaser i formater som SQL), nøgle/værdi-baserede formater (fx JSON) eller opmærkningsbaserede formater (fx XML). I DanDIGI-projektet har vi valgt XML (W3C 2008), som er et tekstbaseret format hvori tekststruktur markeres og metadata tilføjes ved hjælp af XML-elementer bestående af start- og sluttags (afgrænset fra tekstindholdet ved tegnene < og >, fx <title>Dommer til klubstævne</title>).

En afgørende grund til at vælge XML er at det er skabt til at repræsentere tekster og understøtter et ikonicitetsprincip på den måde at diverse indholdskategorier kan opmærkes (navngives) hvor de forekommer i forlægget, snarere end at måtte flyttes ind i en mindre ikonisk struktur, fx navngivne kolonner som i CSV- og SQL-formater eller den lige så lidt ikoniske struktur i JSONs nøgle/værdi-format. Det giver i sig selv vigtige fordele: For det første understøttes det ganske krævende manuelle opmærkningsarbejde der er nødvendigt i

[4] CLARIN (*Common Language Resources and Technology Infrastructure*) er en EU-baseret digital platform hvor forskere på tværs af de europæiske lande kan deponere og dele sprogbaserede data samt værktøjer til at behandle og analysere sprogdata (www.clarin.eu).

vores tilfælde (jf. nedenfor), ved at opretholde en en-til-en-relation mellem forlæg og data som gør det nemmere at bevare overblikket. For det andet bevares i selve grunddataene en væsentlig del af den oprindelige struktur der kan udgøre vigtig kontekstinformation i tolkningen af den repræsenterede interaktion. For det tredje bliver det meget ligetil at skabe en visning der lægger sig tæt op ad den oprindelige brugergrænseflade ved at transformere dataene til webformat (HTML) – vel at mærke en visning hvor følsomme eller uønskede datakategorier let kan undertrykkes eller pseudonymiseres.

Opmærkningen kan valideres ud fra et foruddefineret XML-skema således at det sikres at dokumentet overholder en bestemt struktur. Dette reducerer fejl og sikrer konsistens når data udveksles mellem forskellige it-systemer. Ligeledes vigtigt i forhold til principperne om *reusability* og *interoperability* er det at XML-opmærkningen er selvbeskrivende da dataenes struktur og indholdskategorier er eksplicitet gennem de tags (mærker) der bruges i selve dokumentet.

Der findes en nyligt udgivet standard for XML-opmærkning af data fra sociale medier udviklet af *The Text Encoding Initiative Consortium* (TEI-Consortium 2025), mere specifikt retningslinjerne for opmærkning af *Computer-mediated Communication*, som trådte i kraft i 2024 efter et årelangt forsknings- og udviklingsarbejde (Beißwenger et al. 2012; Beißwenger & Lüngen 2020). Dette arbejde stiller vi os på skuldrene af ved at holde os til det inventar af XML-tags der specificeres i standarden, hvorved vores opmærkning udover at være selvbeskrivende i den ovennævnte forstand også arver et ekstra lag af kategoriinformation fra TEI's omfattende beskrivelsesapparat. I det følgende taler vi derfor også om TEI XML snarere end bare XML.

Vi vil i afsnit 4 nedenfor komme nærmere ind på den konkrete opmærkning af vores data, men først vil vi beskrive de data som indgår i korpusset.

3. Datagrundlag

3.1. Kilder og datatyper

Som det fremgår af tabel 1 nedenfor, stammer teksterne i DanDIGI-korpusset fra tre hovedkilder: Vi udnytter for det første meget store datamængder fra Reddit og Twitter som er stillet til rådighed via

henholdsvis udtræk fra Reddits API[5] og Danish Gigaword-projektet (Derczynski et al. 2021).

For det andet udnytter vi data indsamlet i forbindelse med forskningsprojekter tilknyttet Sprogforandringscentret: Projekt hverdagssprogning (Madsen et al. 2016), Dialekt i periferien (Maegaard et al. 2020), og SoMeFamily-projektet (Hansen & Stæhr 2021). Disse data består af interaktioner på Facebook, Instagram og Messenger, og deltagerne er folkeskole- og gymnasieelever, og i nogle tilfælde også deres forældre, fra København, Hirtshals, Bylderup og Nexø.

Medietype	**Platform**	**Status**	**Størrelse (tokens)**	**Periode**	**Kilde**
Diskussions-forum	Reddit	Offentlig	67 mio.	2014-2023	Academic Torrents (Pushshift.io)
	Heste-Nettet	Offentlig	338 mio.	2000-2024	DanDIGI[6]

[5] Vores Redditdata kommer oprindeligt fra Reddits API (datawebsite), men vi har indtil videre hentet dataene via forskellige tredjeparter. Vores data er således downloadet fra Academic Torrents (https://academictorrents.com/; Lo & Cohen 2015) på det følgende link: https://academictorrents.com/details/56aa49f9653ba545f48df2e33679f014d2829c10 (som er offentliggjort i dette opslag på Reddit: https://www.reddit.com/r/pushshift/comments/1akrhg3/comment/kq994t4/). Dataene er oprindeligt høstet dels af pushshift.io, et projekt der systematisk høster data fra Reddits API (Baumgartner et al. 2020), dels af Redditbrugeren RaiderBDev, der også tilgængeliggør data fra Reddits API via GitHub-kontoen ArthurHeitmann. Academictorrents.com-linket indeholder/giver adgang til de 40.000 største subreddits på Reddit – herunder et beskedent antal danske subreddits hvorfra vi har udvalgt et antal til at indgå i korpusset ud fra nogle simple kriterier: Dels selvfølgelig at de er tilgængelige i det nævnte udtræk (jf. korpussets opportunistiske karakter), dels at teksten overvejende er dansk, og dels at der er mange opslag med over 10 kommentarer.

[6] Heste-Nettet-data foreligger også som en del af Gigaword-korpusset, men i en form uden metadata og interaktionsstruktur. Da det fortsat er juridisk og teknisk muligt at høste data fra Heste-Nettet, valgte vi derfor at gøre det. I DanDIGI-korpusset indgår således indholdet fra samtlige fora under "Ryttersnak" på www.heste-nettet.dk som det forelå primo 2024.

Social netværks-side	Twitter	Offentlig	21 mio.	2019-2020	Gigaword ("General discussions")
	Facebook	Semi-offentlig	1.5 mio.	2011-2019	Sprogforandrings-centret
	Instagram	Semi-offentlig	ca. 150.000	2015-2018	Sprogforandrings-centret
Besked-tjeneste	Messenger	Privat	1.5 mio.	2015-2019	Sprogforandrings-centret

Tabel 1: DanDIGI-projektets data. Bemærk at tallene er omtrentlige. De vil først blive helt præcise når korpusset er færdigopmærket, og det endelige antal af tokens kan optælles maskinelt.

Data repræsenterer således tilsammen seks forskellige platforme med ret forskellige affordances og brugerprofiler og dækker en periode på godt 20 år. Det er ud fra tabel 1 åbenlyst at korpusset er skævt sammensat på den måde at data fra Reddit, Twitter og især Heste-Nettet fylder meget i forhold til Facebook, Instagram og Messenger; men det skal der naturligvis kunne tages højde for ved at man kan søge i de forskellige delkorpora hver for sig eller sammenholde resultaterne med metadata vedrørende platform, produktionstidspunkt osv.

3.2. Formater og metadata

En væsentlig problemstilling i forbindelse med etableringen af korpusset er at data i udgangspunktet foreligger i meget forskellige formater og med metadata i meget forskelligt omfang.

Dataene fra Twitter, Heste-Nettet og Reddit har den store fordel at de foreligger i et veldefineret format som i meget vidt omfang kan viderebearbejdes automatisk. De kan således på trods af deres store omfang inkluderes i korpusset med relativt lille ressourceforbrug. For Twitterdelens vedkommende er en væsentlig mangel til gengæld at rensningen af data mht. at fjerne ikke-brugergenereret tekst og at opnå anonymisering er foregået på en måde så der ikke er nogen metadata mht. hvem der har skrevet hvad, hvornår det er skrevet, og hvilke posts som evt. er reaktioner på hinanden. Det er ikke særlig tilfredsstillende for et korpus som vores, men for Twitterdata er det p.t. den eneste mulighed vi har da indsamling af data efter Twitter er blevet til

X, nærmest er blevet umuliggjort. Heste-Nettet- og Reddit-dataene er derimod indsamlet så interaktionsstrukturen kan genskabes, ligesom vi har oplysninger om hvilke brugere der har skrevet hvad, og på hvilket tidspunkt. Vi har dog ingen oplysninger om brugerne ud over deres brugernavn.

Dataene fra projekterne under Sprogforandringscentret er derimod indsamlet i forbindelse med etnografisk feltarbejde hvilket betyder at vi har mange metadata om såvel deltagerne som konteksten. For en del af deltagerne har vi også lydoptagelser hvoraf en del allerede er transskriberet og inkluderet i LANCHART-korpusset.

En væsentlig ulempe ved dataene fra Sprogforandringscentret er til gengæld at de i udgangspunktet foreligger som screenshots og andre billedfiler hvilket gør at de kun kan behandles semiautomatisk. At få denne del af data opmærket med indholdskategorier i TEI XML er derfor klart den største del af arbejdet med korpusetableringen.

4. Strukturering af data i TEI XML-format

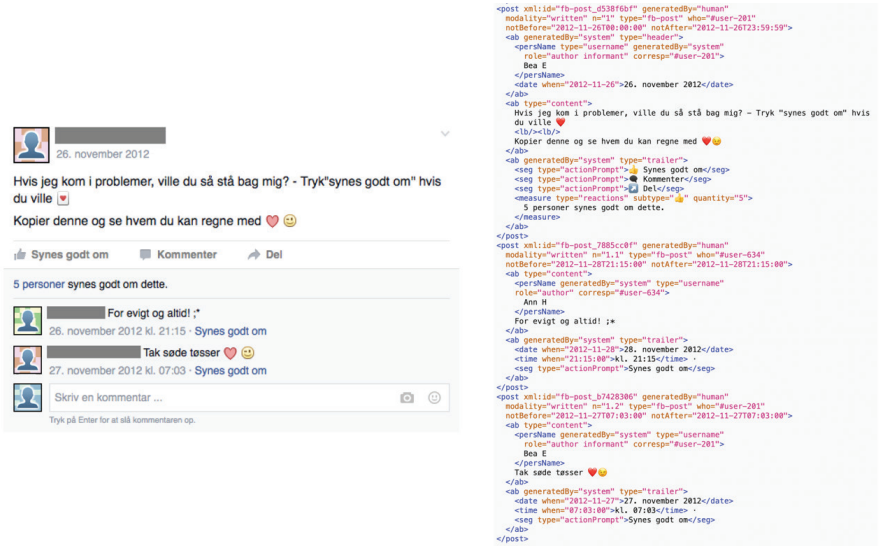
I det følgende vil vi først beskrive endemålet for opmærkningen af vores data, nemlig det valgte TEI XML-format. Herefter vil vi kort beskrive de procedurer der er anvendt for at få de forskellige data opmærket. Vi beskæftiger os i det følgende udelukkende med den *informationsbevarende* opmærkning af korpusset. Korpusset bliver også tilføjet *informationsberigende*, dvs. analytisk, opmærkning med information på ordniveau om lemma, ordklasse og bøjningsform, syntaktisk funktion mm., men det falder uden for denne artikels rammer at komme ind på dette.

4.1. XML-strukturen

I figur 3, 4 og 5 ses tre korte tråde af opslag på hhv. Facebook og Heste-Nettet i form af skærmklip med den dertil svarende TEI XML-opmærkning vist til højre. En tråd er i denne sammenhæng en struktur bestående af et oprindeligt opslag med evt. tilhørende kommentaropslag. De enkelte opslag opmærkes ved hjælp af post-elementet med attributter med metadata om typen af opslag (fx `type="fb-post"`), hvor opslaget forekommer i trådens hierarkistruktur (vha. hierarkisk nummerering a la "1", "1.1", "1.2" osv.) (fx `n="1"`), hvilken bruger der har slået det op (fx `who="#user-11103334ae43"`), og hvornår (fx `notBefore="2010-12-`

`24T00:00:00",notAfter="2010-12-24T23:59:59").` Vi erstatter brugernavne med pseudonyme bruger-id'er, jf. afsnittet nedenfor om pseudonymisering.

Figur 3. Facebookopslag fra 2010 med dertil svarende TEI XML-repræsentation



Figur 4. Facebookopslag fra 2012 med dertil svarende TEI XML-repræsentation.

Figur 5. Opslag fra Heste-Nettet fra 2016 med dertil svarende TEI XML-repræsentation.

Indenfor denne ramme, der er ens for data fra alle platformene, findes opslagets indhold i form af hhv. system- og brugergenererede sektioner. Her er der til gengæld betydelig strukturel variation mellem platformene. Diverse indholdskategorier går igen, fx brugerens brugernavn (`<persName type="username" ...>`) og datoen for opslaget (`<date ...>`), men i forskellig rækkefølge: brugernavnet først eller 'midt' i opslaget (mellem titlen og resten af opslagsindholdet i Heste-Nettets tilfælde); datoen før eller efter opslagsindholdet. Her viser XML-formatet sin styrke idet selve strukturen ikke behøver bære al kategoriinformation – det meste fremgår af navnene på XML-elementerne og deres attributter. Det viser sig dog at der er en vis overordnet struktur der går igen på tværs af de forskellige platforme, nemlig en opdeling i opslagshoved, -indhold og -fod. Opslagshovedet og -foden indeholder systemgenereret information såsom brugernavn, opslagstidspunkt og reaktioner, mens indholdssektionen indeholder brugerens eget indlæg. Dette har vi valgt at afspejle i vores opmærkning vha. sektioner indlejret i ab-elementer (*ab* for "anonymous block")

af typen ”header”, ”content”, ”trailer”[7] – især for at kunne medtage denne sektionsopdeling i vores kontekstvisning af dataene.

I de viste eksempler varierer rækkefølgen af sektioner som vist i tabel 2.

Facebook 2010	Facebook 2012	Heste-Nettet 2016
• Header: Brugernavn. • Content: Indlæg. • Trailer: Dato m.m. • Trailer: Reaktioner.	• Header: Brugernavn + dato + lokation. • Content: Indlæg. • Trailer: Reaktionsknapper. • Trailer: Reaktioner.	• Header: Titel + brugernavn + dato. • Content: Indlæg. • Trailer: Reaktioner + antal visninger samt emne (ikke synligt i skærmklippet).

Tabel 2: Strukturen i de tre opslagseksempler (relevant for det oprindelige opslag, dvs. det første post-element i tråden – i kommentaropslagene er rækkefølgen/strukturen igen en anden).

En detalje man særligt bør lægge mærke til, er at der ikke er nogen garanti for at en sektion indeholder ren system- hhv. brugergenereret tekst. I Heste-Nettet-opslagene indeholder opslagshovedet således en brugergenereret titel efterfulgt af systemgenererede forfatter- og datooplysninger. Det fremgår tydeligt af eksemplet at titlen kan være en integreret del af brugerens tekst, og den skal derfor selvfølgelig med som løbende tekst i korpusset, i modsætning til den øvrige systemgenererede tekst, der har karakter af metadata. Dette sikres ved at titlen er forsynet med attributten ”generatedBy” med værdien ”human”.

De mere eller mindre subtile strukturelle forskelle til trods giver XML-opmærkningen således mulighed for at identificere de samme indholdskategorier i de forskellige data mhp. at gøre dem søgbare på lige fod i korpusset.

Post-elementet (dvs. det enkelte brugeropslag) er den grundlæggende enhed i interaktion på sociale medier; men vores opmærkning skal også rumme metadata på højere niveauer, især proveniensoplysninger på korpus- og subkorpusniveau. TEI XML-standarden er designet til at rumme sådanne oplysninger i et hierarki af korpusser og dokumenter, og DanDIGI-korpusset må siges at udnytte disse mulig-

[7] Dertil kommer typen opslagsindskud (et ab-element af typen ″insert″) som indeholder systemgenereret information i en sektion mellem indholdssektioner (ikke relevant i de viste eksempler).

heder til fulde. Vores data er således indlejret i det følgende strukturelle hierarki, med relevante metadata på hvert niveau:

- TEI-korpus: DanDIGI-korpusset
    - TEI-subkorpora (Sociale medietyper): sociale netværkssider, diskussionsfora, beskedtjenester.
        - TEI-subkorpora (Platforme): Facebook, Twitter, Instagram, Heste-Nettet, Reddit, Messenger.
            - TEI-subkorpora (Kilde/projektlokation): Periferiprojektet-Bornholm, SoMe Family osv.
                - TEI-dokumenter (Forum/gruppe/profil): Bruger X' Facebookprofil, Gruppe Y's Messengergruppe osv.
                    - tekst-elementer (Hvert tekst-element indeholder en tråd af opslag).
                        - post-elementer de enkelte opslag

4.2. Opmærkningsprocedurerne

Visse af vores data kan som nævnt processeres automatisk. Det gælder dog ikke for Facebook-, Messenger- og Instagramdataene der foreligger som billedfiler og derfor til dels må håndteres manuelt. Kort fortalt er proceduren for Facebook-, Messenger- og Instagramdataene at der udføres en OCR-genkendelse af teksten, en automatisk udtrækning af evt. billeder i opslagene som siden kan vises i vores kontekstvisning, evt. i anonymiseret form, samt en automatisk råparsing (dvs. tentativ strukturgenkendelse og opmærkning) af den givne datafil, alt sammen vha. en række til formålet udviklede Pythonscripts, hvorefter den resulterende opmærkning tilrettes manuelt. Det manuelle arbejde er tids- og opmærksomhedskrævende, så det skal understøttes på enhver tænkelig måde – herunder ved at følge det ovennævnte ikonicitetsprincip.

4.3. Pseudonymisering

En væsentlig del af kodningen af vores data angår navne og billeder der kan være direkte personhenførbare. Da vi samtidig ikke kan garantere at dataene ikke også er personfølsomme visse steder (det er jo ikke utænkeligt at informanter omtaler hinandens politiske tilhørsforhold,

seksualitet eller andre personfølsomme oplysninger), er det såvel af etiske hensyn som af hensyn til GDPR-reglerne (EU 2016) afgørende at personoplysningerne håndteres forsvarligt. Det gør vi dels ved at sørge for at dataene kun anvendes indenfor rammerne af den såkaldte forskningshjemmel (Databeskyttelsesloven § 10), dels ved at navne pseudonymiseres[8] og billeder sløres.

Ikke desto mindre forsøger vi at bevare facetter af de oprindelige navne og billeder – dette for at kunne tilbyde en kontekstvisning (se afsnit 5) hvor billeder og navne så vidt muligt kalkerer de oprindelige, inkl. signaler om køn, alder, etnicitet m.m., for at bevare så mange potentielt betydningsbærende træk fra de oprindelige data som muligt. Som et fiktivt eksempel ville personnavnet *Tórunn Heinesen* således fx kunne pseudonymiseres som *Oddvør Herdal*. Dette opnår vi ved at lade AI-modeller indsætte pseudonymer der ligner de oprindelige navne distributionelt, og AI-genererede billeder der ligner de oprindelige billeder visuelt (i øvrigt alt sammen udført lokalt med open source-modeller for at undgå at sende data via nettet til store udbydere af AI-services). Denne proces fører det for vidt at gøre nærmere rede for her; det vil vi gøre i anden sammenhæng.

I selve XML-dataene erstattes navne af frene pladsholderpseudonymer fra et begrænset inventar (Ann, Bea, Cia, Dea osv.)[9] samt af anonyme unikke brugerkoder af formen *user-c1eb079a7edf* placeret i opmærkningen af de givne navne. Således ville den fiktive Tórunn Heinesen fx kunne repræsenteres sådan:

```
<persName type="username" role="author"
corresp="#user-c1eb079a7edf">Dea A</persName>
```

Selve brugerkoden bruges til at henvise til en selvstændig database med alle navnedata (rigtigt navn, brugernavne fra de forskellige platforme samt diverse pseudonymer) for alle personer i korpusset. På denne måde vil alle navne i korpusset være grundigt pseudonymiseret

[8] Man taler om *pseudonymisering* snarere end *anonymisering* når de oprindelige navnedata stadig kan genskabes, selv hvis det kun er af en snæver kreds af projektmedarbejdere (EU 2016, artikel 4.5).

[9] Pladsholderpseudonymerne, som ikke er unikke for hver bruger, bruges primært for at bevare antallet af tokens fra de oprindelige data (i modsætning til fx at erstatte et navn på to tokens som *Tórunn Heinesen* med den ét token lange brugerkode).

selv hvis de rå XML-data deles med andre forskere. Samtidig bliver det meget mere ligetil at vedligeholde visningspseudonymerne (og fx ændre et pseudonym hvis det af den ene eller anden grund ikke er passende) da de er registreret ét og kun ét sted i navnedatabasen hvortil der så refereres vha. brugerkoderne.

Hvad billederne angår, er de placeret direkte i XML-dataene i form af kodestrenge (såkaldt base64-encodede billeddata) – men i den AI-slørede version således at der heller ikke her ligger data der er personhenførbare.

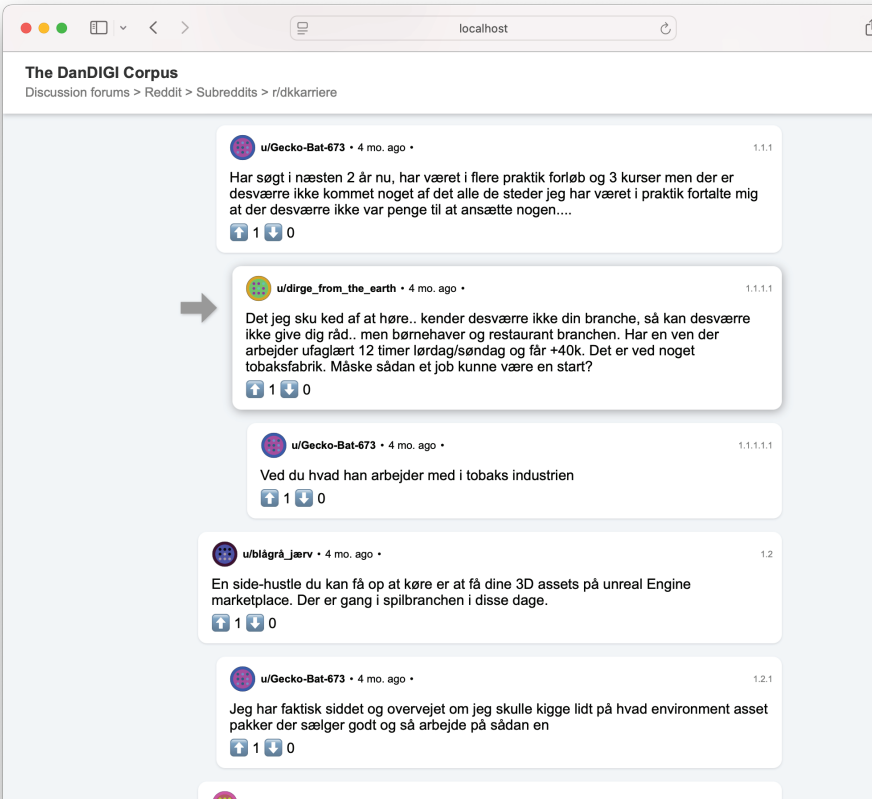
5. Konvertering til andre formater

XML-formatet er vores basale format i projektet, dvs. det som vi tilføjer informationer i, og det som er lagrings- og udvekslingsformatet for eftertiden. På basis af dette velbeskrevne format med samtlige informationer kan man så generere andre formater til brug i specifikke sammenhænge. I DanDIGI arbejder vi med to sådanne afledte formater – dels VRT (VeRticalized Text), dels en kontekstvisning i HTML. Begge formater har som formål at danne grundlag for den nutidige forskningsmæssige udnyttelse af korpusset, både for forskere internt på Sprogforandringscentret og for forskere fra andre forskningsinstitutioner. De danner således grundlag for den måde korpusset vil blive tilgængeliggjort for forskere, dvs. via et webinterface (med adgangsbegrænsning).

VRT-formatet er et simpelt format som kan importeres i den korpusinfrastruktur vi i forvejen har kørende på LANCHART-centret, nemlig korpusdatabasen Corpus Workbench og søgeværktøjet Korp. Korp understøtter en række korpuslingvistiske metoder, dvs. diverse søgemuligheder og resultatvisninger samt forskellige former for statistiske analyser og sammenligninger (Borin et al. 2012; Diderichsen & Jensen 2023).

Kontekstvisningen i HTML bliver et supplement til Korps standardkonkordansvisning. Her udnytter vi ikoniciteten i vores strukturopmærkning til at lave en visning der ligner den oprindelige visning – med den væsentlige fordel ift. en egentlig faksimilevisning (dvs. en visning med de oprindelige billedfiler) at vi har kontrol over hvilke oplysningstyper vi vil vise, hvilke vi ikke vil vise, og hvilke

vi vil vise i en pseudonymiseret form. Dette er afgørende i forhold til muligheden for at stille data til rådighed for andre forskere. Figur 6 viser et eksempel fra vores aktuelle prototype for denne visning.



Figur 6. Udsnit af prototype af kontekstvisning, her med data fra Reddit.

6. Nye muligheder for forskning

Etableringen af DanDIGI-korpusset åbner en lang række af nye muligheder for forskning i sprog og interaktion på sociale medier i en dansk kontekst.

For det første giver korpusset mulighed for sammenligninger af interaktion og sprogbrug på tværs af de medietyper og platforme som er repræsenteret i korpusset, også over tid. Dette åbner fx for studier

i hvordan sprogbrugen afspejler mediets affordances, og hvordan konventionerne for anvendelsen af sproglige udtryk udvikler sig over tid.

For det andet giver integrationen af DanDIGI-korpusset i Sprogforandringscentrets korpusinfrastruktur mulighed for på en meget direkte måde at sammenligne hverdagsskriftsprog med talesprog. Dette skyldes at det her bliver muligt at kombinere søgninger i DanDIGI-korpusset med søgninger i det store LANCHART-korpus med dansk talesprog (Diderichsen & Jensen 2023; Sprogforandringscentret u.å.). Både mht. medieplatforme og tale- vs. skriftsprog kan sammenligningen i et vist omfang ske på individplan idet vi for en del af informanterne i LANCHART-dataene har sprogbrug fra flere sociale medier og tale- såvel som skriftsprogsdata.

For det tredje giver det at korpussets sprogbrugsdata både er velstrukturerede, omfangsrige OG informationsrige (qua korpussets metadata og bevarede interaktionsstruktur), mulighed for i langt højere grad end hidtil at kombinere kvalitative og kvantitative metoder i studiet af digitalt medieret interaktion. Man vil således fx kunne fortolke kvantitative mønstre i sprogbrugen fundet gennem datadrevne analyser i Korp med nærsproglige analyser af data vist i kontekstvisningen, eller omvendt undersøge generaliserbarheden af udsagn baseret på kvalitative analyser via korpuslingvistiske metoder.

7. Konklusion

DanDIGI-projektet etablerer et omfattende, bredt sammensat og velstruktureret korpus med data fra sociale medier. Det er vores forhåbning at vi gennem valget af TEI XML-formatet og den tilstræbte overholdelse af FAIR-principperne sikrer dets langtidsholdbarhed og anvendelighed på tværs af forskellige forskningsinteresser. Det er også vores forhåbning at korpusset vil blive udvidet med yderligere data i fremtiden ved at andre forskere bidrager med data, og at projektets erfaringer med opmærkning af samt søge- og visningsformater for digitalt medieret interaktion kan være til nytte for andre der arbejder med eller påtænker at indsamle nye data fra sociale medier.

DanDIGI-korpusset vil efter planen være tilgængeligt for forskere i løbet af 2026 via Sprogforandringscentrets installation af Korp på lanchartkorp.ku.dk.

Litteratur

Baumgartner, Jason, Zannettou, Savvas, Keegan, Brian, Squire, Megan & Blackburn, Jeremy (2020). The Pushshift Reddit Dataset. *Proceedings of the international AAAI conference on web and social media*, 14, 830-839.

Beißwenger, Michael, Ermakova, Maria, Geyken, Alexander, Lemnitzer, Lothar & Storrer, Angelika (2012). A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative*, 3. https://doi.org/10.4000/jtei.476.

Beißwenger, Michael & Lüngen, Harald (2020). CMC-core: a schema for the representation of CMC corpora in TEI. *Corpus*, 20. https://doi.org/10.4000/corpus.4553.

Borin, Lars, Forsberg, Markus & Roxendal, Johan Korp – the corpus infrastructure of Språkbanken *Proceedings of LREC 2012* (s. 474-478). ELRA.

Derczynski, Leon, Ciosici, Manuel R., Baglini, Rebekah, Christiansen, Morten H., Dalsgaard, Jacob Aarup, Fusaroli, Riccardo, Henrichsen, Peter Juel, Hvingelby, Rasmus, Kirkedal, Andreas, Kjeldsen, Alex Speed, Ladefoged, Claus, Nielsen, Finn Årup, Madsen, Jens, Petersen, Malte Lau, Rystrøm, Jonathan Hvithamar & Varab, Daniel (2021). The Danish Gigaword Corpus. I Simon Dobnik & Lilja Øvrelid (red.), *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (s. 413-421). Linköping University Electronic Press.

Diderichsen, Philip & Jensen, Torben Juel (2023). Samtaler i korpusformat: Repræsentation af talesprog i LANCHARTs korpus-infrastruktur. *Nordlyd*, 47(2), 77-89. https://doi.org/10.7557/12.7084.

DSL (u.å.). *KorpusDK*. Det Danske Sprog- og Litteraturselskab. Hentet 03.01.2025 fra https://ordnet.dk/korpusdk.

EU (2016). *EUROPA-PARLAMENTETS OG RÅDETS FORORDNING (EU) 2016/679 af 27. april 2016 om beskyttelse af fysiske personer i forbindelse med behandling af personoplysninger og om fri udveksling af sådanne oplysninger og om ophævelse af direktiv 95/46/EF (generel forordning om databeskyttelse)*. https://eur-lex.europa.eu/legal-content/DA/TXT/?uri=CELEX:32016R0679

Frey, Jennifer-Carmen, König, Alexander, Stemle, Egon, Falaise, Achille, Fišer, Darja & Lüngen, Harald (2020). The FAIR Index of CMC Corpora. I Julien Longhi & Claudia Marinica (red.), *CMC Corpora through the prism of digital humanities* (s. 127-144). Paris: L'Harmattan.

Gibson, James Jerome (1977). The theory of affordances. I Robert Shaw & Bransford (red.), *Perceiving, acting, and knowing: toward an ecological psychology* (s. 67-82). Hillsdale, N. J.: Lawrence Erlbaum.

Hansen, Marianne Haugaard & Stæhr, Andreas Candefors (2021). Sproglige generationsforskelle på de sociale medier. *NyS, Nydanske Sprogstudier,* 59, 113-156. https://doi.org/10.7146/nys.v1i59.121811.

Hutchby, I. (2001). Technologies, Texts and Affordances. *Sociology*, 35(2), 441-446.

Jensen, Klaus Bruhn (2013). Computermedieret kommunikation. I Gunhild Agger, Agnete Nørgård Kristensen, Per Jauert & Kim Schrøder (red.), *Medie- og kommunikationsleksikon*. Frederiksberg: Samfundslitteratur. https://medieogkommunikationsleksikon.dk/computermedieret-kommunikation-2/.

Lo, Henry Z. & Cohen, Joseph Paul (2016). Academic Torrents: Scalable Data Distribution *Neural Information Processing Systems 2015* (s. 1-2).

Lomborg, Stine (2025). Sociale medier. I Gunhild Agger, Agnete Nørgård Kristensen, Per Jauert & Kim Schrøder (red.), *Medie- og kommunikationsleksikon*. Frederiksberg: Samfundslitteratur. https://medieogkommunikationsleksikon.dk/sociale-medier-2/.

Madsen, Lian Malai, Karrebæk, Martha Sif & Møller, Janus Spindler (red.) (2016). *Everyday Languaging: Collaborative Research on the Language Use of Children and Youth*. Berlin, München, Boston: De Gruyter Mouton. DOI: 10.1515/9781614514800.

Maegaard, Marie, Monka, Malene, Køhler Mortensen, Kristine & Candefors Stæhr, Andreas (2020) *Standardization as Sociolinguistic Change: A Transversal Study of Three Traditional Dialect Areas*. Milton: Routledge. DOI: 10.4324/9780429467486.

McEnery, Tony & Hardie, Andrew (2012) *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

Sprogforandringscentret (u.å.). *LANCHART-korpusset*. Sprogforandringscentret. Hentet 03.01.2025 fra https://dgcss.hum.ku.dk/online-ressourcer/lanchart-korpusset/.

TEI-Consortium (2025, 24.01.2025). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.9.0*. TEI Consortium. Hentet 01.04.2025 fra http://www.tei-c.org/Guidelines/P5/.

W3C (2008, 26.11.2008). *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. World Wide Web Consortium. Hentet 02.01.2025 fra https://www.w3.org/TR/REC-xml/.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18.