

Corpus of Danish Digitally-mediated Interaction (DanDIGI) er opbygget af forskellige datakilder, der kræver manuel behandling i forskelligt omfang for at komme fra rådata til TEI XML – det fælles format bag de forskellige visninger af dataene (korpusvisning, kontekstvisning). Arbejdsgangen har syv trin:

Trin 1 Facebook- og Messenger-data foreligger som brugerdonerede skærbilleder. Hestenettet-, Reddit- og Twitter-data foreligger i forskellige maskinlæsbare formater og kan behandles automatisk.



Fig. 1 Originalt skærbilledudsnit (redigeret).

Trin 2 DanDIGI-udviklet automatisk XML-strukturering (Facebook og Messenger kræver også OCR).

Trin 3 DanDIGI-designet, AI-drevet pseudonymisering: Rigtige navne gemmes i en fortrolig database og erstattes af platformefterlignende pseudonymer.



Fig. 2 AI-drevet pseudonymisering.

Trin 4 DanDIGI-designet billed-'pseudonymisering' gennem AI-imitation af billedindhold.



Fig. 3 AI-drevet billed-'pseudonymisering'.

Trin 5 Kun for Facebook og Messenger: validering + manuel korrektion. XML håndkurateres; emojis, OCR-tastefejl, datoer osv. rettes, hvilket gør dette trin meget arbejdskrævende.



Fig. 4 Håndkurateret, pseudonymiseret TEI XML.

Trin 6 Grammatisk annotering af ord- og sætningsstruktur ved hjælp af DanGram-taggeren.

Trin 7 DanDIGI-udviklede automatiske transformationer til 1) en korpusvisning, der muliggør sproglig søgning, eksport og kvantitative analyser, og 2) pseudonymiserede kontekstvisninger, der nøje efterligner de originale platforme til at se fund i kontekst og understøtte kvalitative interaktionsstudier.

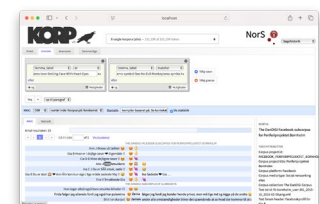


Fig. 5 Korpusvisning.

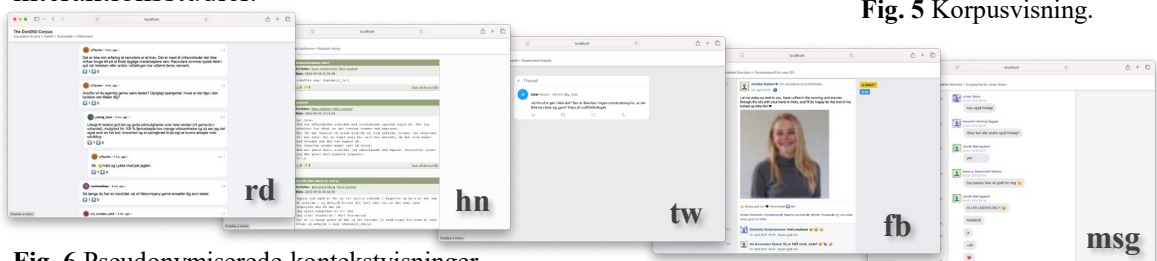


Fig. 6 Pseudonymiserede kontekstvisninger.