Foreløbige analyser af samtaledata fra Børns Vilkår

WP1 af 'Understanding Childrens' Discourses on Well-Being'

Kristian G. Kjelmann Matias K. Appel
Josefine M. Christensen

2026-02-01

Indholdsfortegnelse

Baggrund

Denne rapport samler arbejdet i WP1 af pilotstudiet "Understanding Childrens' Discourses on Well-Being", 2025. Arbejdspakken havde til formål at foretage indledende analyser af samtaledata fra Børns Vilkår. Følgende gennemgår data, metoder og foreløbige resultater. Analyserne har leveret indsigter, der udgjorde grundlag for senere forskningsansøgninger. Der blev foretaget to analyser: (1) en foreløbig emneanalyse på tværs af kanaler og (2) analyse af semantiske relationer i sprogbruget. Resultaterne peger på et distinkt semantisk sprogbrug knyttet til bestemte kernepersoner i børns omgangskreds (forældre, venner, skolelærere) samt deres primære sociale arenaer (fx skole, hjem). Der er indikationer på variationer mellem køn, men robustheden er endnu ikke tilstrækkelig til at drage stærke konklusioner.

Metoden generelt

Arbejdet består af to parallelle analyser. For det første en emneanalyse, hvor der undersøges dominerende tematikker i det, som børn kontakter Børnetelefonen om. Hertil er det undersøgt, om tematikker varierer på tværs af platforme. For det andet en undersøgelse af "semantiske relationer" i det, som der bliver talt om, som er udledt via såkaldte "word embeddings"-modeller. Denne analyse involverer at træne modeller baseret på ord, der optræder i kontekst, for at udforske ligheder og forskelle i, hvordan bestemte roller, personer og arenaer bliver italesat. Her har der været særligt opmærksomhed på forskelle mellem piger og drenge. De foreløbige resultater indikerer, at modellen kan fange forskelle, men at resultaterne i øjeblikket er sårbare over for små ændringer i data og derfor ikke er stabile nok til fortolkning.

Om data

Datagrundlaget består af samtaledata fra chat og sms, brevkassebeskeder samt fem fokusgruppeinterviews. Fokusgruppeinterviews er inddraget for at få indikation af, om det, som børn italesætter, når de kontakter Børnetelefonen, også fremgår i kvalitative opfølgninger. De beskrivende statistikker af Tabel 1 samtale- og brevkassedata. Data er filtreret til brugere i alderen 10-16 år. Såkaldte "faste brugere" og beskeder fra rådgiver er frasorteret. Samtaler er sammenlagt på tværs af beskeder, så hver samtale fremstår som én samlet tekst, der kan analyseres ensartet. Der er arbejdet med samtale- og brevkasse data fra Q4 2024 og Q1 2025.

Fordeling af data

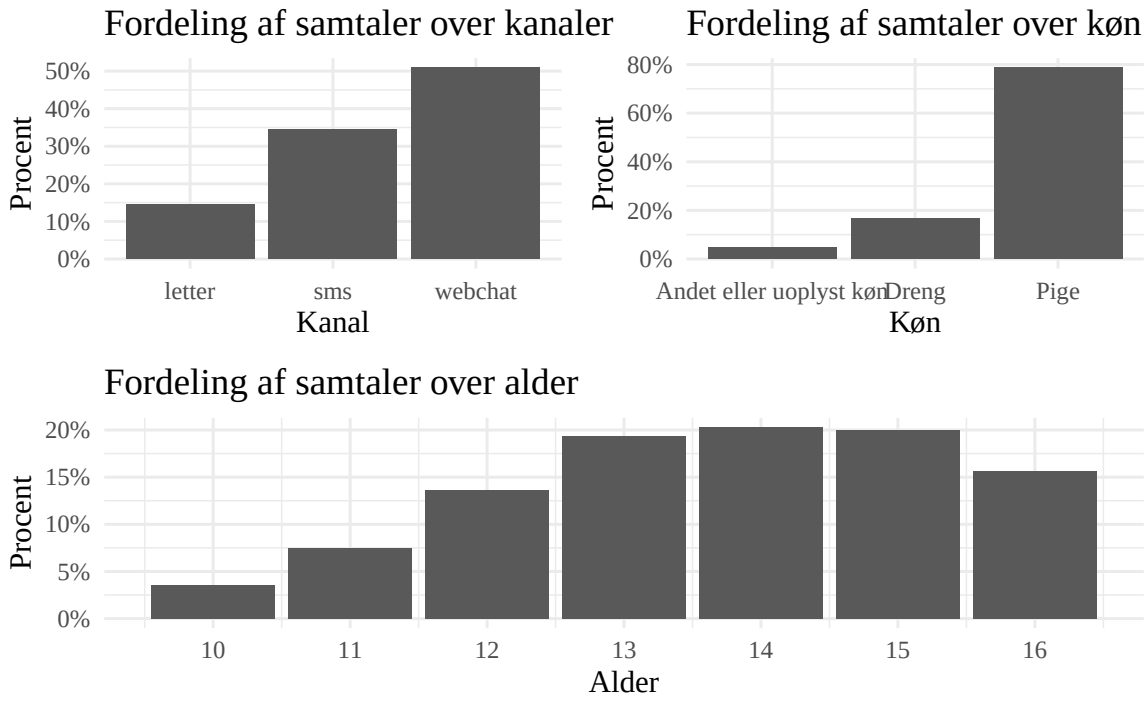
For at forstå hvem materialet repræsenterer, beskriver vi populationen via fordeling af samtaler på kanaler, køn og alder. Figur 1 viser et samlet overblik over fordeling i brug af kanal,

Tabel 1: Beskrivende statistikker

Beskrivende statistikker

Antal chats	11.700
Total antal tokens	3.457.670
Gns. antal tokens per chat	295,53
Dato for første chat	2024-10-01
Dato for sidste chat	2025-03-27

kønsfordeling samt aldersfordeling. Fordelingen er tydeligt skævvredet ang. køn med ca. 80% piger. Webchat er den mest populære platform (50 %), og det er aldersgruppen 13-15, som er mest tilstedeværende i materialet.



Figur 1: Fordeling af samtaler efter kanal, køn og alder.

Emneanalyse

Baggrund og metode

Emneanalysen bygger på samtale-, brevkasse- og interviewdata, som er inddelt i lige store tekststykker på cirka 150 tegn. Data er filtreret til aldersgruppen 10-16 år og renset for faste brugere. For samtaledata betyder det, at vi kun bruger brugerens beskeder og frasorterer meget korte beskeder som "hej" og "nej". Hvert tekststykke konverteres til en numerisk repræsentation ved hjælp af en eksisterende sprogmodel ("sentence transformer"), hvor hvert stykke tekst tildeles en numerisk og sammenlignelig repræsentation. Derefter anvender vi klyngeanalyse (HDBSCAN) til at gruppere tekststykker efter indhold, hvilket gav omkring 120 emner. Disse blev efterfølgende manuelt inspiceret og samlet på baggrund af en kvalitativ vurdering, hvilket reducerede antallet til 54 emner.

Nøgleord på tværs af data

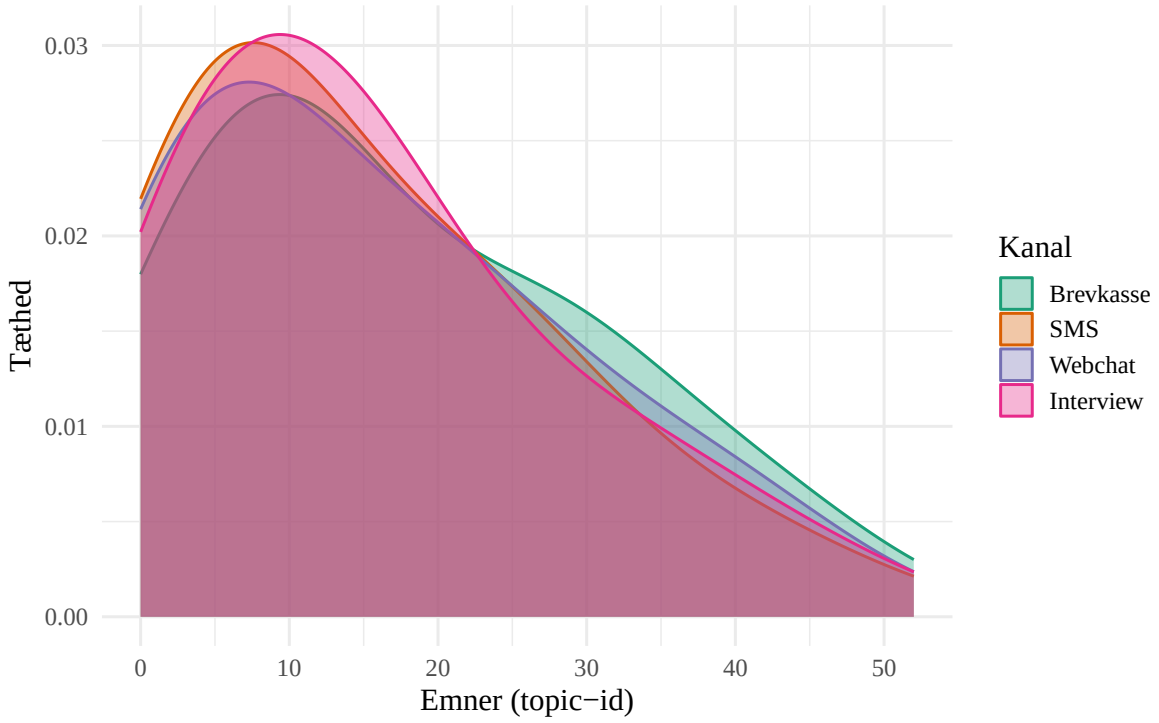
For at få et første indtryk af emner i materiale, udledes de væsentligste nøgleord for hver af de 54 emner. Meget almindelige ord og tekniske støjord fjernes. Figur 2 viser centrale nøgleord for emnerne, hvor størrelsen angiver forekomst i materialet.

Figur 2: Nøgleordsprofil på tværs af emner. Størrelse angiver forekomst.

Fordeling af emner på tværs af kanaler

Figur 3 viser fordelingen af alle emner på tværs af kanaler. Dog er et enkelt emne udeladt, da det alene udgjort 40 % af interviewdata. Dette emne er derfor frasorteret for bedre at kunne sammenligne de resterende emner. Resultaterne indikerer for det første en tydelig prævalens af bestemte emner. Derudover viser figuren, at fordelingen af emner ligner hinanden på tværs af kanaler (med undtagelse af interviewdata, der som nævnt var domineret af ét emne, som her er ekskluderet).

Generelt giver denne foreløbige model et fint indblik i dominerende tematikker, men yderligere arbejde er nødvendigt for bedre at kunne forstå, hvad emnerne reelt afspejler, hvordan de er forskellige, og om der er demografiske forskelle i, hvilke emner der bliver italesat.



Figur 3: Fordeling af emner på tværs af kanaler (alle emner).

Semantiske relationer i sprogbrug

Data brugt til modellerne

Semantiske relationer er udledt via embeddingmodeller. Embeddingmodellerne er trænet udelukkende på chat- og sms-data fra brugere i alderen 10-16 år, og faste brugere er frasorteret. Kun beskeder fra brugeren indgår. Brevkassedata er ekskluderet ud fra en betragtning om, at samtaleformen og skrivestilen varierer markant fra chat- og sms.

Teksterne er tokeniseret, dvs. opdelt i enkeltord, hvor tegnsætning fjernes, ord konverteres til ordstammer, og stopord udelades. Der er trænet tre modeller: én på alt materiale, én på drenge og én på piger. Personer, som identificerer sig som andet køn end dreng eller pige, er ikke inkluderet, da de udgjorde en meget lille gruppe. Tabel 2 viser, hvor mange tokens der indgår før og efter filtrering samt for de kønsspecifikke modeller.

Tabel 2: Data til embedding modeller.

Data til embedding modeller

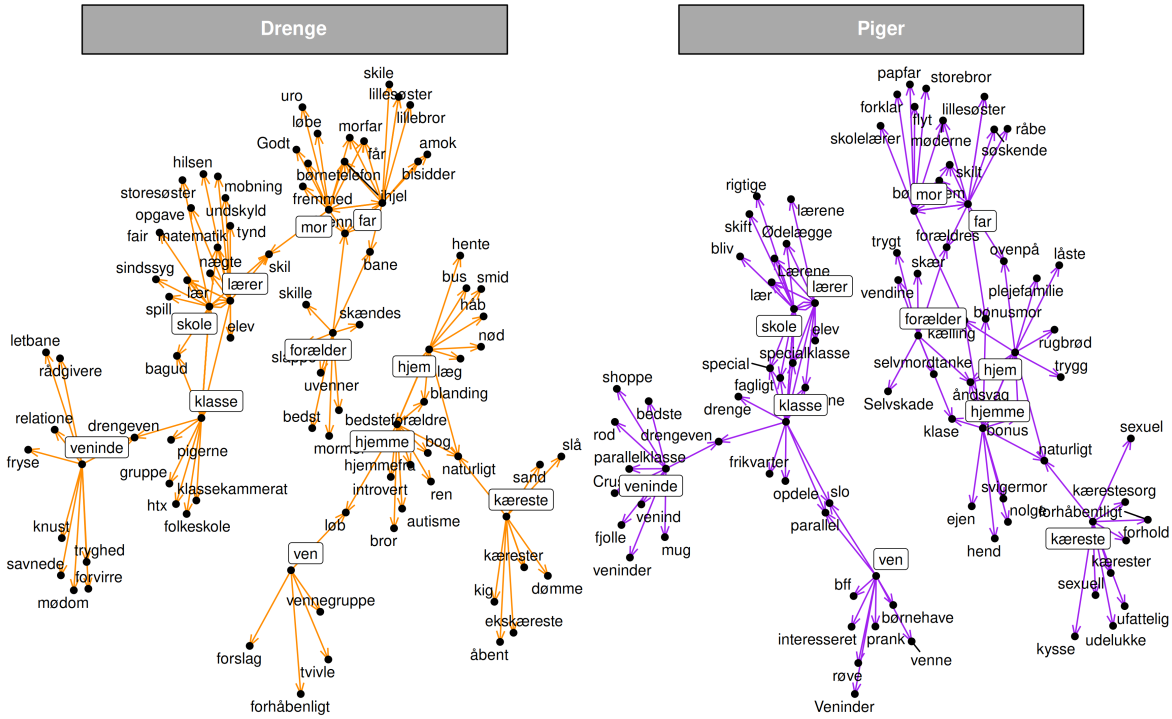
Total antal tokens	3.457.670
Antal tokens efter filtrering	416.013
Antal tokens - Drenge	69.814
Antal tokens - Piger	331.998

Metodisk baggrund

Embeddingmodellerne forsøger for hvert ord at forudsige, hvilke ord der optræder i nærheden, og bruger en kontekst på 10 ord på hver side. Resultatet er en såkaldt "vægtvektor" på 100 tal for hvert ord, som afspejler, hvilke kontekster ordet typisk optræder i. Ord med lignende vægte tolkes som semantisk beslægtede. For at undersøge semantiske relationer, der der udledt de 10 mest semantisk beslægtede nøgleord for en række forud definerede nøgleord, som vurderes relevant ift. målgruppens sociale liv: forældre, venner, skole og hjem.

Kønsforskelle i semantiske relationer

Vi sammenligner placeringen af nøgleord i de kønsspecifikke modeller for at se, om relationer og nærliggende ord varierer mellem drenge og piger. Figur 4 er baseret på de 10 mest lignende ord til "mor, far, lærer, skole, klasse, veninde, forælder, hjem, hjemme, ven, kæreste" i begge kønsspecifikke modeller. Den giver foreløbige indsigter i centrale associationer til sociale roller og arenaer, og peger på tydelige kønsforskelle.



Figur 4: Kønsforskelle i semantiske relationer baseret på top-10 mest lignende ord.

Modelvalidering

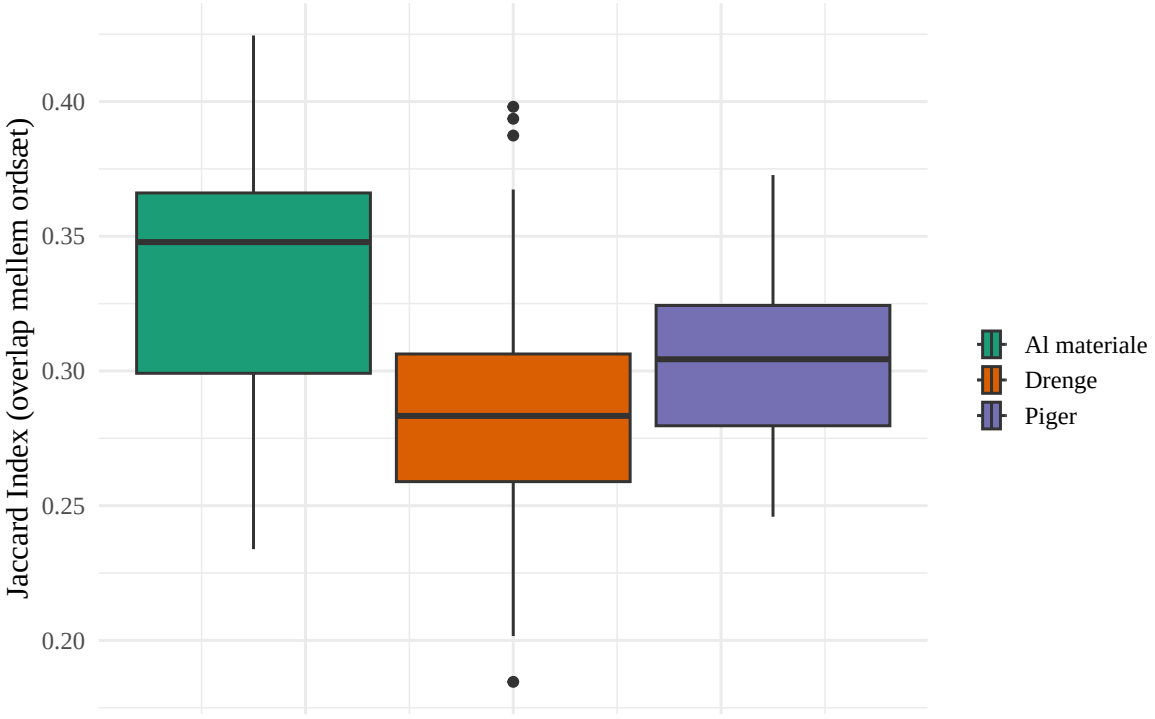
Baggrund

Semantiske modeller evalueres normalt ud fra, om de indfanger de "forventede" relationer. Dog er formålet med denne analyse netop at få indblik i mulige semantiske relationer, som før har været overset, hvorfor relevant valideringsdata ikke findes. I stedet er robustheden af resultater undersøgt ved krydsvalidering. Der trænes 10 modeller, hvor 10 procent af data udelades i hver træning. Hver gang udledes de fem mest lignende ord for de forud definerede nøgleord. Overlap mellem disse ordlister måles med Jaccard-indekset, hvor 0 betyder intet overlap, og 1 betyder fuldt overlap. Højere værdier afspejler mere stabilitet på tværs af træninger, mens lavere værdier indikerer større variation.

Resultat af validering

Resultaterne viser generelt lave gennemsnitlige Jaccard-indekser (Figur 5), hvilket indikerer, at de fundne ord varierer meget mellem modeller, når små dele af data udelades. Det peger

på, at de nuværende modeller endnu ikke er stabile nok til sikker fortolkning.



Figur 5: Jaccard-indeks for robusthed på tværs af modeller.

Videre arbejde

I det videre arbejde fokuseres på embeddingmodellerne. De nuværende modeller er statiske embeddings, hvor hvert ord har én fast repræsentation uanset kontekst. Dynamiske (kontekstuelle) embeddings kan i stedet tilpasse repræsentationen til den konkrete sætning og fange variation i betydning på tværs af brugssituationer. Det kan forbedre både stabilitet og tolkbarhed, men kræver mere data og beregningskraft. Samtidig er der allerede indikationer af kønnede forskelle i semantiske relationer, som det er relevant at følge op med mere robuste og kontekstuelle modeller. Som vist er resultaterne dog ikke robuste, og indikerer behov for yderligere træningsdata, eller en opsætning, der udnytter andre tilgængelige sprogressourcer (som fx modeller, der i forvejen er trænet til "typiske" semantiske relationer på dansk).