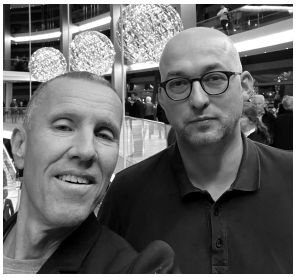
Ny teknologi i en gammel verden



Kunstig intelligens og oldhebraiske tekster

Af postdoc, ph.d. Martijn Naaijer og lektor, ph.d. Martin Ehrensvärd

I løbet af de seneste årtier er det blevet klart at afskrivere blev ved med at redigere i og udvide de hebraiske bibeltekster relativt langt op i tiden. På denne måde fastholdt man teksternes relevans. En bedre forståelse af denne proces er afgørende for fortolkningen af teksterne.

I et to-årigt forskningssamarbejde mellem Det Teologiske Fakultet og Datalogisk Institut ved Københavns Universitet undersøger vi Det Gamle Testamentes tekst- og sproghistorie med udgangspunkt i bibelmanuskripter fundet blandt Dødehavsrullerne. Studiet er baseret på kunstig intelligens og maskinlæring, et område som er i rivende udvikling. Der findes endnu ingen metoder til at foretage en sådan undersøgelse, så en vigtig del af projektet går ud på metodeudvikling. Projektet startede 1. oktober 2021.

Kunstig intelligens

Kunstig intelligens er blevet væsentlig mere intelligent i løbet af det seneste årti. Dette kan ses på fx Google-søgninger eller på Google Translate. Vores smartphones er også et godt eksempel. Udviklingen sker trinvist og kommer snigende. Men hvis man sammenligner nutidens telefoner og tjenester med dem for ti år siden, kan man se en forbløffende udvikling. Dette projekt ville heller ikke have været muligt før for ganske nylig.

Der er sket store forbedringer på tre fronter. For det første har nye metoder inden for maskinlæring vist sig at være meget effektive. For det andet har hardwareudvikling nået et stadie, hvor omfattende beregninger er bredt tilgængelige uden at man skal vente for længe på at computeren får regnet færdig. De kan endda køres gratis på fx Googles servere. For det tredje, hvilket er det vigtigste for dette projekt, er mange af oldtidens tekster nu digitaliseret og frit tilgængelige online på fx GitHub.com.

Kunstig intelligens er et vidt begreb. En vigtig, konkret form for kunstig intelligens er maskinlæring. Ved hjælp af maskinlæring kan vi nu analysere sproglige udtryk, der er både komplekse og almindelige, hvilket er krævende hvis menneskelige hjerner skal gøre det uden computerhjælp. Vi kan skabe et glimrende overblik over materialet, og vi kan måske gøre det med mindre menneskelig bias end det var muligt før, idet vi nemmere med hårde facts kan underbygge eller afvise mere intuitive formodninger. For tiden undersøger vi fx James Barrs teori

om at man oftere stavede hebraiske ord defektivt (altså uden konsonanttegn brugt som vokalbogstaver) når ordene har et affiks (som fx et pronominalsuffiks).

Semitiske filologer og lingvister har traditionelt forsøgt at absorbere så mange sproglige data som menneskeligt muligt. Herigennem udvikler de deres intuition, og det hjælper dem til at rekonstruere sprogenes historie og til at forstå nye tekster, når de graves frem. Vigtige, gamle semitiske tekster dukker op hvert år. Dette er en god og nyttig fremgangsmåde. Men mennesker har deres begrænsninger.

Nye metoder

Metoden "Natural Language Processing" (NLP) søger at forstå og generere naturligt sprog. Den blev grundlagt allerede i 1950'erne, og optimismen var stor da en IBM mainframe-computer som proof of concept oversatte en række russiske sætninger til engelsk i januar 1954, kendt som Georgetown–IBM-eksperimentet. Det viste sig dog at være et meget svært problem at løse ordentligt. Tidligere brugere af Google Translate vil vide, hvor dårlig en oversætter Google var (sammenlignet med dygtige mennesker) indtil for nylig. Men hvis man ikke har brugt Google Translate i løbet af de seneste år, vil man blive overrasket over, hvor meget bedre det fungerer i dag. Googles neurale netværk producerer sætning efter sætning af ofte godt engelsk, dansk, fransk osv., og ofte – om end ikke altid – korrekt oversat.

Nye modeller og teknikker, såsom Deep Learning, har været afgørende i denne udvikling. I dette projekt anvender vi disse teknikker på oldhebraisk. Vi har desuden fået afgørende inspiration fra bioinformatik, hvor kolossale genetiske strenge analyseres. Disse strenge minder om almindeligt sprog.

Problemet med hebraisk

Det Gamle Testamente udgør til dels tekstgrundlaget for tre verdensreligioner, men dets tidlige tilblivelse er stadig gådefuld. Tekstens udvikling studeres af tekstkritikere, og sprogets udvikling af historiske lingvister. Det er for det meste adskilte forskningsfelter, men for at få et fuldstændigt billede af teksternes historie – fordi vi ofte har flere eller mange forskellige versioner af den samme hebraiske tekst – skal de integreres. Det overordnede formål med projektet er at udvikle en metode til at integrere disse felter ved hjælp af nye maskinlæringsteknikker så vi får et mere komplet billede af udviklingen af både tekst og sprog end det er muligt i øjeblikket.

De hebraiske bibeltekster er undergået en lang proces med affattelse, redaktion og transmission. De ældste overlevende manuskripter er de versioner af teksterne som Dødehavssamfundet producerede og brugte. Men teksterne havde allerede udviklet sig i århundreder, hvor de blev udvidet og ændret løbende. Vi ved at der er dukket mange variationer op under transmissionen af teksten, da tekster ofte blev kopieret ufuldstændigt.

Mange forskere er enige om at de bibelske teksters historie og hebraisk sproghistorie er beslægtede problemer, og at man ikke meningsfuldt kan studere det ene uden det andet. I praksis er tekstkritik og historisk lingvistik dog stadig to adskilte forskningsområder. De fleste undersøgelser af hebraisk sproghistorie er baseret på et middelaldermanuskript, Co-

dex Leningradensis (L, dateret til 1008-1009 e.Kr.). Sprogforskere der bruger dette manuskript som deres hovedkilde, hævder at dets sprog er blevet bevaret gennem mange århundreder.

Men konsensus blandt tekstkritikere er at den teksttradition der findes i Codex Leningradensis, blot er en af mange traditioner, og at den ikke bør gives den fremtrædende plads som sprogforskere har tildelt den. Det er fordi tekstvariation er mere tilfældig end ofte antaget. Dette er et stort stridspunkt inden for forskningsfeltet eftersom mange hebraiske sproghistorikere er uenige med tekstkritikerne i dette.

Ved at undersøge hvordan teksterne i manuskripter grupperer sig, og hvor (in) konsistent sproglig variation er mellem og inden for manuskripter, kan man opnå en bedre forståelse af, hvordan teksttransmission og sproghistorie tog sig ud.

Projektets spørgsmål og mål

Vores forskerhold er tværfagligt og omfatter Martin Ehrensvärd, specialist i hebraisk, Anders Søgaard, specialist i maskinlæring, og Martijn Naaijer, hvis ph.d.-afhandling bygger bro mellem disse to felter.

Det overordnede spørgsmål som vi stiller os selv, kan formuleres sådan: Hvordan kan maskinlæring kombineret med sekvensanalyseteknikker udviklet inden for bioinformatik anvendes som et analytisk værktøj til at studere de bibelske dødehavsruller og bidrage til (1) studiet af de hebraiske bibeltekster flydende natur, og (2) beskrivelsen af hebraisk sproghistorie?

Forskningen har to hovedformål. Det første mål er at etablere metoder til at analysere teksternes flydende natur og deres sproglige variation. Der findes ingen gode metoder til dette, så vi vil udvikle nogle ved hjælp af såkaldte Graph Neural Networks. Det andet mål er at udvikle og anvende metoderne til bredt at analysere tekstlige og sproglige traditioner, som også vil kunne anvendes på andre antikke sprog.

Indsigter i den store Esajasrulle

Vi arbejder foreløbig på den store Esajasrulle fra Qumran. Det er en næsten komplet udgave af Esajas' Bog, og det er samme version af bogen som vi kender fra Det Gamle Testamente. Visse bøger som fx Jeremias har vi overleveret i meget forskellige versioner. Den udgave af Jeremias som oversætterne af Septuaginta brugte, var en hel del kortere, og en del af materialet står i forskellig rækkefølge. Den (for os) alternative udgave af Jeremias fandtes også i Qumran og blev brugt side om side med vores udgave.

Esajas' Bog havde dødehavssamfundet hele 22 udgaver af (mindst), men altså 22 afskrifter af samme version. De fleste af de manuskripter vi har fundet, er dog stærkt fragmentariske. At de afspejler samme grundversion, er dog bestemt ikke ensbetydende med at udgaverne var ens. I masser af detaljer varierer manuskripterne, også i forhold til vores udgave i Codex Leningradensis. Det er alt fra ordvalg til stavning.

Langt de almindeligste varianter i stavning forekommer ved de såkaldte matres lektionis, altså læsemødre, dvs. de førnævnte konsonanttegn der bruges til at indikere vokaler. Jo ældre en tekst, desto færre matres. Nogle af de tekstformer der kendes fra Qumran, er typologisk senere

end den vi kender fra Codex Leningradensis, grundlaget for de fleste moderne oversættelser. Så disse manuskripter er fulde af matres. Det gælder også den store Esajasrulle, eller i hvert fald første halvdel.

Man har nemlig længe været opmærksom på muligheden for at anden halvdel af manuskriptet blev skrevet af en anden afskriver end den første. Dels er der et lille ophold efter første halvdel, og dels er stavningen i anden halvdel ret forskellig og minder mere om Codex Leningradensis' mere sparsomme brug af matres.

Håndskriften i begge halvdele ser ens ud, men ved hjælp af kunstig intelligens har forskere i Holland påpeget systematiske forskelle. I vores projekt har vi spurgt os selv, om det mon var almindelig praksis at man deltes om opgaven med at afskrive så stort et manuskript som Esajas. Kun tre af de øvrige Esajas-manuskripter indeholder nok tekst til en meningsfuld undersøgelse, og vores undersøgelse viser at brugen af matres i første og anden halvdel er konstant, så det tyder ikke på at være tilfældet.

Vi har også arbejdet på nogle datasæt som vi vil offentliggøre på GitHub. Dels den samaritanske udgave af Første Mosebog, dels Esajasbogens ca. 3000 substantiver hvor vi har fjernet affikser, altså præpositioner, bestemte artikler, nominalendelser, pronominalsuffikser etc. Substantiverne i Esajas har indtil videre udgjort grundlaget for vores arbejde med at sammenligne brugen af matres i de to halvdele og i de øvrige Esajas-manuskripter.