

Om digitalisering af spørgelistespørgsmål

Line Sandbæk Geertsen & Stine Vejbæk Rathkjen

Peter Skautrup Centret for Jysk Dialektforskning, Aarhus Universitet

Indledning

Jysk Ordbog, som hører hjemme på Peter Skautrup Centret for Jysk Dialektforskning på Aarhus Universitet, er bygget på en samling af materiale og kilder, som hovedsageligt er fysiske, ikke digitale. Således står millioner af sedler i fyldte kartoteksskuffer, og tusindvis af spørgelistebesvarelser står opstillet på hylder, hvilket næsten er en selvfølge, da arbejdet med at indsamle materiale om og på jysk dialekt begyndte ved centrets grundlæggelse i 1932. Spørgelistesamlingen blev påbegyndt i 1949, da den første avisspørgeliste om husdyrhold blev udgivet i forskellige lokalaviser. Siden da er samlingen vokset til en imponerende størrelse med 193 forskellige spørgelister. Langt de fleste af disse er udsendt i perioden 1949-2014 (til cirka 500-700 meddelere), mens enkelte spørgelister kommer fra andre kilder, f.eks. Udvalg for Folkemål (nu Center for Dialektforskning), Københavns Universitet, som står bag tre spørgelister i samlingen.

I midt-september påbegyndte vi, som en del af vores praktik på Peter Skautrup Centret, digitalisering af selve spørgsmålene i spørgelisterne, sådan som de blev sendt ud til meddelerne. Processen og resultatet heraf vil vi beskrive i denne rapport.

Spørgelisternes form og brug

Spørgelistespørgsmålene er arkiveret i seks bøger og en mappe. Bøgerne dækker over spørgelisterne B til DØ, samt “særlige spørgelister” og “unavngivne lister”. I mappen findes de geografiske lister, som er blevet udsendt til specifikke egne i Jylland med særtilpassede spørgsmål. I alt består samlingen af godt 4600 spørgsmål fordelt på 193 spørgelister.

De fleste spørgelister hører til del B-Å og AB-DØ og har en relativt ensartet udformning: en række spørgsmål i venstre kolonne og felter til besvarelse i højre kolonne. Spørgelisterne er typisk 3-6 sider lange og har 10-50 spørgsmål og underspørgsmål. På hver spørgeliste findes også et felt på enten første eller sidste side, hvor meddelelsen bliver bedt om at redegøre for sin alder, sit hjemsgn, og hvor mange år meddelelsen har boet i sit sogn.

Figur 1 viser et typisk eksempel på en side fra en spørgeliste. Her findes spørgsmål om ords udtale, eksistens og varianter. Spørgsmål 36 og 38 er delt i to underspørgsmål. Spørgsmålene inkluderer typisk behjælpelige forslag til svar, skrevet på en lydlig gengivelse af dialekt, uden den lydskrift, som Jysk Ordbog bruger. Dog med undtagelse af længdemarkøren <·>.

Et betydeligt antal spørgelister har imidlertid et atypisk format. Dette omfatter primært de tidligste og de 'særlige' spørgelister, f.eks. Husdyrspørgelisen eller Mindre spørgeliste 1, se figur 2 og 3.

De besvarede spørgelister er især blevet brugt til at gøre det klart, hvor visse betydninger og udtaler af ord er attesteret i Jyllands mange sogne. Dette visualiseres bl.a. i Jysk Ordbog ved hjælp af udbredelseskort. Nogle af spørgelistebevisererne er også tilføjet til seddelsamlingen, hvorved de indgår i redaktionen af ordbogen. At digitalisere selve besvarelserne er også et igangværende projekt, som kræver manuel indtastning af tekst grundet den flerartede form i besvarelserne og af hensyn til GDPR. En nærmere beskrivelse af dette kan findes i Holm (2024).

Hvordan

Hovedformålet med digitaliseringen af spørgelistespørgsmålene er først og fremmest at gøre dem søgbare samt at gøre materialet tilgængeligt, uden at man behøver at møde fysisk op på centret. Online tilgængelighed kan nemt løses ved scanning og upload, mens ønsket om søgbarhed kræver et andet format. Her faldt valget på Excel. Fordelene ved dette valg er mange: Excel er udbredt, letlæseligt, og xlsx-data kan nemt kopieres og bruges i andre programmer og formater.

	Spl. CL 4
34. Hvordan udtales i Deres dialekt ordet <u>vant</u> i sætningen "Du er vist godt <u>vant</u> "? (wänn, wond, vån, van, wänt eller anderledes).	
35. Kendes i Deres dialekt ordet <u>vant</u> (= forpint eller bange for), brugt f.eks. i sætningen: "Jeg er så <u>vant</u> i mine tænder"? Hvordan udtales ordet? (wont, wänt, want, wäntj, vant eller anderledes). Giv om muligt andre eksempler på ordets brug.	
36 a) Hvordan udtales i Deres dialekt ordet <u>vejrlig</u> i sætningen: "Sikke en (eller et) grim(t) <u>vejrlig</u> "? (weerle, wejrlaw, wejrlow, wøjle, væle eller anderledes)?	
b) Kan ordet anvendes i betydningen "Ånde, luft", f.eks. i sætningen: "Jeg kunne næsten ikke få mit <u>vejrlig</u> igen".	
37. Hvordan udtales i Deres dialekt ordet at <u>vifte</u> i sætningen "Bladene begyndte at <u>vifte</u> "? (hwøww, wøwt, wøwwt, hwøwwt, vøjjt, vøft eller anderledes).	
38 a) Kendes i Deres dialekt ordet <u>vik</u> , brugt f.eks. i sætningen "Det er en bitte <u>vik</u> en", og hvad betyder ordet?	
b) Kan ordet i Deres dialekt bruges i forbindelsen "at være <u>vik</u> for noget", og hvad betyder det?	
39. Kendes i Deres dialekt udtrykket: "Han er den <u>vildeste</u> af dem" med betydningen: "den bedste" eller "den dygtigste"? Giv eksempel på ordets brug.	
40. Hedder det <u>aparte</u> eller <u>apartig</u> i Deres dialekt? Angiv udtalen (aparte, aparter, aparti, aparto eller anderledes). Sæt streg under den del af ordet, som har tryk.	
41. Hvordan udtales i Deres dialekt ordet en <u>vogn</u> ? (uwn, own, wuwn, wåwn, wukwen, wuun, un, won eller anderledes).	
42. Hvordan udtales i Deres dialekt ordet at <u>vogte</u> (om hyrde o.lign.)? (ut, wut, wot, vøwwt eller anderledes).	
43. Kendes i Deres dialekt ordet en <u>vol</u> brugt om en gren, som endnu er så tynd, at den kan bøjes? Hvordan udtales ordet? (wul, wol, wull, woll eller anderledes).	
44. Kendes i Deres dialekt udtrykket: "Det kan <u>volde</u> sig" med betydningen "Det kan ske, indtræffe"? Hvordan udtales <u>volde</u> i denne forbindelse? (woll, wåld, wull, ull, voll eller anderledes).	

Figur 1: Spørgeliste CL, side 4. Ikke udfyldt.

C. Svinet.	
1. Almindelig Betegnelse for Arten: (Grise, Svin, Svinhoveder)	håndskrift + kort aug. 1995
2. Alder: Uden Henblik til Køn. 0-2 å 3 Mdr. (Gris, Putteris, hette Gris)	
HS C 2 a 2 å 3-4 å 5 (Polt, Ungsvin, Drivesvin, Stantling, Stuntling)	håndskrift + kort aug. 1995
4 å 5- (Gris)	
Hundgr. 0-2 å 3 (Ornegris, Orninggris, Galtgris, Hangris)	
HS C 2 b 2 å 3-4 å 5 (Orne, Orning, Orningpolt, Galtgris)	håndskrift + kort sept. 1995
4 å 5- (Orne)	
Hundgr. 0-2 å 3 Mdr. (Sogris)	
2 å 3-4 å 5 (Sogris, Sopolt, So)	
HS C 2 c 4 å 5- (Sogris, So, Grisso, Leves)	håndskrift + kort sept. 1995
Kastreret. 0-2 å 3 (Galtgris, Galt, Gris)	
2 å 3-4 å 5 (Galt, Polt, Stantling)	
HS C 2 d 4 å 5 (Galt)	håndskrift + kort sept. 1995
Kastreret som voksen. 4 å 5- (Ornegalt, Gilding, Seg)	
3. Hundeyret, der første Gang har Unge: (Galt, Grisso, Ungso, So, Førstefagss)	håndskrift + kort sept. 1995
4. Udtryk for, at Hunnen er brunslig: (vred o. l.)	håndskrift + kort maj 1989

Figure 2: Husdyrspørgelisten, side 6.

Vi valgte at gå til denne digitalisering på to forskellige måder, som beskrives nedenfor. Stine anvendte AI-chatbotter, som både kan aflæse tekst fra billeder og formatere det aflæste i et Excel-ark. Line, som har haft undervisning i programmering og computerlingvistik, valgte at kode sit eget program for at gøre det samme.

Mar. Spt. 3
mnd. Spt. 1
1955

se *hmm* mappes 10. sept.

Jysk ordbogs mindre spørgeliste 1.

I jyske dialekter findes forskellig udtale af den lyd, der svarer til rigsmålets d (i for eksempel gade, mad). Den kan enten udtales som d eller r eller g eller j, eller den kan helt falde bort. Vi kan altså fra Jylland hente følgende forskellige former af for eksempel ordet mad: mad, mag, mag, maj eller ma. Instituttet ville være Dem taknemmelig, hvis De ville angive udtalen af ordene i denne spørgeliste med særlig henblik på den omtalte lyd, der er understreget med .

Hvis De bor i et grænseområde mellem to udtaler, det vil sige: kender De en anden udtale af den adspurgte lyd fra et nærliggende sogn, beder vi Dem i nederste højre hjørne omtrentligt angive, hvorfra formen eller lyden stammer.

1) (at) bide	Ordene i denne række behøver ikke særlig anførelse i nederste højre hjørne.
4) (jeg) bijer	
5) (en) gryde <i>håndhids. i holt juni 89</i>	
6) (flere) gryder	25) (at) lade (noget være)
7) (at) skyde	26) lad (være)
8) (je-)skyder	27) (han) lod
9) (at) vide	28) bedre
10) (jeg) væl	29) køgel <i>håndhids. i holt</i>
11) (at) lade	30) tredje
12) (jeg) læger	31) noget
13) (at) slige	32) meget
14) (jeg) sliger	33) gå ind ad doren
15) kød	34) følges ad
16) brød <i>håndhids. i holt nov. 89</i>	35) lan lavege
17) fløde <i>håndhids. i holt sept. 89</i>	36) han har lavet
18) hedder	37) broget - <i>håndhids. i holt feb. 88</i>
19) (jeg) gad	38) kroget - <i>håndhids. i holt</i>
20) (en) gade	39) måned <i>håndhids. i holt</i>
21) (flere) gader	40) fremmed - <i>håndhids. i holt marts 85</i>
22) (en) klud	41) tag 'et rolig - <i>håndhids. i holt</i>
23) (en) knude	
24) (flere) knuder	
25) lod	
26) Knud	
Meddelelser:	Fra (sogn eller by, kan høres udtalen med d / r / j / g / bortfald. (Understreg den lyd - eller de lyde - De kender).
stedt den	
sogn	

Figur 3: Mindre spørgeliste 1 fra 1955.

Stines digitalisering med brug af AI-værktøjer

I forsøget med digitalisering af spørgelisterne virkede det oplagt at undersøge muligheden for brug af generative AI-værktøjer som f.eks. Gemini, CoPilot og ChatGPT. Da jeg, Stine, personligt har størst erfaring med at bruge ChatGPT, var det denne, jeg valgte til opgaven. Før arbejdet kunne begynde, var der dog nogle potentielle problematikker, jeg skulle tage stilling til.

Ifølge AU's vejledning om brug af generativ AI bør man altid være varsom med, hvilket materiale man uploader, da man aldrig kan være helt sikker på, hvordan materialet opbevares og/eller bruges af GAI-værktøjet ([Aarhus Universitet 2025](#)). Før digitaliseringsarbejdet kunne igangsættes, var det derfor vigtigt først at undersøge, om der i spørgelisterne indgik nogen form for personfølsomme oplysninger eller andre former for sårbart materiale. Dette var ikke tilfældet. Især fordi formålet med digitaliseringen er at gøre materialet offentligt tilgængeligt på sigt, var dette vigtigt at undersøge.

I arbejdet med GAI-værktøjer som ChatGPT er det afgørende for processen at starte med den rette prompt. En prompt kan forstås som en "instruks eller forespørgsel til [...] en kunstig intelligens om at udføre en bestemt handling" (Det Danske Sprog- og Litteraturselskab 2025). Fra tidligere erfaringer med GAI ved jeg, at det bedste resultat ofte opnås ved at starte med en relativt udførlig forklaring af opgaven og det ønskede resultat. Min første prompt til ChatGPT lød derfor således:

(a) "Jeg har nogle lister med spørgsmål. De er skrevet på papir, og jeg vil gerne have skriften digitaliseret og sat ind i et excel-ark. Jeg vil gerne have tre kolonner i excel-arket. En med nummeret på spørgsmålet, en med spørgsmålet og en tom kolonne til højre. Hvis jeg giver dig et billede med teksten, kan du så læse, hvad der står?"

Det første resultat fra ChatGPT var umiddelbart imponerende. Tal og tekst var opdelt og placeret korrekt, og der var indsat en tom kolonne som ønsket. ChatGPT havde endda aflæst og indsat en overskrift på spørgelisten, i dette tilfælde Liste AB, samt gengivet kursivering. Ved nærmere inspektion var resultatet dog ikke helt så imponerende som først antaget. Ca. en fjerdedel af dialekteksemplerne viste sig at være det pure opspind. En påstand jeg naturligvis gerne ville have kunnet dokumentere med et billede. Desværre bliver store filer ikke gemt ret længe. Jeg har forsøgt at genskabe det første produkt ved at give ChatGPT samme prompt og samme spørgeliste både på en anden konto og på programmets gratisversion, men med to vidt forskellige resultater. Ingen af dem lig det første resultat jeg fik. Ved de nye forsøg er der ikke så meget tale om påfund fra chatbottens side, men mere

om mangelfuld aflæsning, især af vokaler. Det, at arbejdsprocessen ikke kan genskabes, understreger et generelt problem ved brug af GAI i academia, hvor ethvert givent resultat altid bør kunne efterprøves (Fink 2021:13).

Når dette er sagt, var de indledende problemer med påfund hurtige at løse. Ved at give ChatGPT besked om, at eksemplerne (markeret med parentes i spørgelisten) er skrevet på dialekt, og at den derfor skal se bort fra almindelig retskrivning, ikke må digte, og skal markere det, den ikke kan læse, i stedet for at fylde huller ud, blev problemet faktisk løst. Med den nye prompt begyndte ChatGPT endda at medtage længdeprikker <·> og stødtegn <’>.

Overordnet set var ChatGPT et godt redskab til at løse opgaven, for med et par simple rettelser i min prompt blev spørgelisterne læst hurtigt og korrekt. Ydermere havde værktøjet ingen problemer med skift i format, manglende tal, håndskrevne tegn oveni teksten, tekst, der overskred tekstfeltet, eller andre potentielle problematikker i de uensartede spørgelister. Uagtet inputtet forblev outputtet det samme. Præcis som jeg havde bedt den om. Se figur 4. Resultaterne skulle selvfølgelig korrekturlæses for at sikre, at der ikke var fejl. ChatGPT havde som oftest problemer med vokaler, hvor én vokal blev skiftet ud med en anden, men jo flere spørgelister den blev præsenteret for, desto færre fejl var der at finde i dens output. Allerede efter 7-10 spørgelister fandt jeg kun mellem 0-3 fejl pr. liste, og da resultatet blev givet i en færdig xlsx-fil, gik digitaliseringen hurtigt.

Nr.	Spørgsmål	Kommentar 
1	Om det at tøjre en ko ved noget meget saftigt græs, således at den kun kunne nå en smal kant ad gangen, siger man i nogle egne: den må do noge ind eller måske: den må kun få en (el. et?) nog. Hvad hedder det på Deres dialekt? Kendes andre ord for det samme begreb? Kendes nog og noge i andre betydninger?	
2	Hvad kaldte man på Deres dialekt en bikube? (bihyve, bikube, bikurv (bikårre)? En bikube med bier kaldtes: bistade (bistå, biståd), bistage (biståg), bistan (bi-stand), bistok? (Skriv, hvis betydningen var anderledes).	
3	Hvad hed en bolle i suppe? (bolle, klimpe, klump, kløse, kumpe, munk). (Var det melbolle el. kødbolle el. begge dele?)	
4	Hvad kalder De gøgespyt? (Den skummende, spytagtige klump, der kan findes på blomster, og som omgiver skumcikadens larve) (blat, fandens spyt, fuglespyt, gøgespyt, heks  t, kukkerspyt, kukkuspyt, kukmands-spyt, troldespyt, trøldespyt, trolakonespyt, trolsmør, tudespytt). Huskes	

Figur 4: Eksempel på besvarelse fra ChatGPT.

Lines selvprogrammerede OCR

Både for at undgå at benytte AI og for at udvikle mine egne færdigheder valgte jeg (Line) at bruge de programmeringsevner, jeg havde tilegnet mig på min kandidat i lingvistik. Jeg undersøgte muligheden for at udføre en OCR-skanning (Optical Character Recognition) i Python. Dette kan relativt nemt gøres ved hjælp pakken *pytesseract*. Denne pakkes hovedfunktion er netop at læse tekst ud fra billeder. Det første skridt i enhver digitalisering er dog at få det fysiske materiale ind i computeren, hvorfor jeg tilbragte en rum tid ved skanneren for at omdanne spørgelisterne til pdf-format.

Som al anden kodning afhænger resultatet af, at inputtet er i et korrekt format, hvilket for *pytesseract* er en billedfil, her PNG, med god billedkvalitet. For at håndtere billeder i Python anvendte jeg den udbredte pakke *OpenCV*. Med *OpenCV* kunne jeg forbehandle billederne for at få dem i den bedst mulige kvalitet til OCR-skanning; jeg konverterede billederne til sort/hvid og fjernede så vidt muligt

de vandrette og lodrette streger, som opdeler spørgsmålene på spørgelisten. Sidstnævnte valgte jeg at gøre, da *pytesseract* læste linjerne som bogstaver – de lodrette linjer blev ofte skannet som *i* eller *l*. Specielt denne del af processen var ikke optimalt udført, og jeg oplevede stadig, at *pytesseract* kunne læse dem som bogstaver, men det lå desværre uden for min formåen at forbedre denne del af koden. Jeg skabte også tre varianter af skanningsfunktionen, da jeg ofte ville skære noget af højre del af billedet fra, da det inkluderede uønsket tekst, som var især gældende for DA-DØ og de områdespecifikke spørgelister, hvor der var håndskrevne noter. Andre spørgelister havde til gengæld tekst på tværs af hele siden. Derfor brugte jeg enten den snævre funktion, som fjernede halvdelen af billedet, den brede, som bevarede hele billedet, eller typisk den mellemste funktion, som skar den højre kvartdel væk.

Jeg afprøvede først dette på en enkelt side af en enkelt spørgeliste, hvor jeg fik et godt resultat. Kun et par bogstaver var fejlskannede, instruks *beskriv* blev til *måskriv*, og der var enkelte forbytninger mellem komma og punktum. Jeg skabte her også den første udgave af den funktion, som konverterer den skannede tekst til en formateret excel-fil, som beskrevet ovenfor. Jeg skrev et regulært udtryk, som tillader en at finde og udvælge stykker fra en længere tekst. Denne skulle finde alle tilfælde af et tal efterfulgt af et punktum, hvorved spørgsmålene kunne opdeles til forskellige rækker i Excel. Sidenhen er dette regulære udtryk blevet forlænget og kompliceret, når der opstod forskelle i, hvordan spørgelisterne formaterede spørgsmålene. Først og fremmest skulle jeg tilføje muligheden for også at opdele spørgsmålene baseret på bogstaver, som adskiller underspørgsmålene, se figur 1. Det regulære udtryk skulle samtidig ikke opfange bogstaver, som reelt var del af spørgsmålet, for eksempel i det hyppigt forekommende *f.eks.* Det regelmæssige udtryk skulle også ændres undervejs, da underspørgsmålene i visse spørgelister, især de områdespecifikke spørgelister, f.eks. SVD, ikke længere markeres med en parentes men med et punktum. Det endelige udtryk fremgår af (b).

(b) `((?:(<=&D)|(<=&b))(<!--)\d{1,2}\b)(?:(<!--el.))|(<!--el.))
 ((?<=[.,\n])[aAbBcCdDeEfF].(?:[?.,])(! ?eks))((?:(<=&D)|(<=&b))
 (<!--)\d{1,2}[.,])(?!d)[.,\n]?([aAbBcCdDeEfF].)(?![?.,])`

Selvom den første side, som jeg afprøvede, gav et godt resultat, gav mange andre spørgelister OCR-skanninger af lavere kvalitet. Noget af det kunne tages højde for ved at 'rette' resultatet, inden jeg arbejdede videre med teksten. Dette blev i høj grad også gjort med regulære udtryk, hvormed jeg kunne fjerne uønskede linjebud og konvertere nogle tegnsamlinger til anden tekst. Der var f.eks. ofte en sammenblanding af *i*, *l*, og *l*, hvilket hindrede den ovenstående opdeling. Men da disse altid opstår i forskellige konstruktioner, kunne jeg erstatte et *l* efter et ikke-bogstav og foran et punktum med *l*. Jeg konstruerede omtrent 50 af sådanne udbytninger, som hver OCR-skanning blev kørt igennem. Mange udbytninger var dog ikke relevante for alle skanninger. Ligeledes blev den første del af spørgelistens første side udelukket, samt meddelerfeltet på første eller sidste side.

For at omdanne det skannede stykke tekst til en excel-fil benyttede jeg python-pakken *pandas*. *Pandas* er praktisk til at arbejde med tabeller og har en indbygget funktion, som konverterer en tabel til en excel-fil. Den var derfor det ligefremme valg til dette digitaliseringsprojekt. Ved hjælp af en let modificeret udgave af det regulære udtryk (b) kunne jeg placere tal, bogstav og selve spørgsmålet ind i de korrekte rækker og kolonner. I enkelte situationer var der dog problemer med at skanne tal og bogstaver korrekt, hvorfor de blev placeret i samme celle som foregående spørgsmål. Dette måtte desværre korrigeres manuelt.

I sidste ende krævede det stadig en vis mængde korrekturlæsning, alt efter kvaliteten af skanningen, før excel-filen var acceptabel. Ikke sjældent blev spørgsmål ikke opdelt, hvis skanningen ikke opfangede spørgsmålets tal. Dette var især et problem, hvor selve spørgelistens tryk var dårligt (figur 5), eller tallet var usynligt (figur 6). Der var ofte artefakter fra de vandrette og lodrette linjer og de håndskrevne noter i spørgelisterne DA-DØ og de særlige spørgelister, som var umulige at undgå helt. Generelt var de særlige spørgelister mere korrekturkrævende, da deres format ikke på samme måde kunne dækkes af min programmering, som var baseret på de regelmæssige spørgelister. Understregninger kunne også være problematiske, da de hindrede korrekt læsning af visse bogstaver og tegnsætning, som det kan ses i figur 7; understregning blev desuden ikke bevaret i OCR-

skanningen. Længdemarkøren <->, som ikke er en del af *pytesseracts* OCR-skanningsmodel, skulle altid placeres manuelt.

3. Hvad kaldes i Deres dialekt planten <u>gyvel</u> (gyvl, gejl, gil)?
4. Hvad kaldes i Deres dialekt frugten af <u>hyben-rosen</u> (hyben, hybtorn, hjo-f eller andet)?
5. Hvilket udtryk bruges i Deres dialekt for at koen føder en kalv (koen kaller, <u>køller</u> , får kalv)?

Figur 5: Spørgsmål 3-5 fra spørgeliste AC.

1. Hvordan udtales i Deres dialekt ordet <u>at bøje</u> (å bøj, å byy, å bøjw, å byw)?
2. Hvad kaldes i Deres dialekt en <u>ås</u> på ploven, en plovås (os, ows, böm, bäm, plovsböm, plovsbäm el.a.)?

Figur 6: Spørgsmål 8-9 fra spørgeliste BI.

5. Hvilke(t) ord bruger man i Deres dialekt for 'overmåde fed, smækfed': <u>blæverfed, bælgfed, bællefed, bælgtyk, bælletyk, kvabset, kvabvorn, pommerfed, pølfed, skjælfed, skrupfed, smadderfed, smækfed</u> eller andre?

Figur 7: Spørgsmål 5 fra spørgeliste AE.

Sammenfatning og sammenligning

I sidste ende var kodningen ikke ligefrem hurtigere eller mindre fejlfri end AI-løsningen, omend fejlene sjældent var så alvorlige som AI'ens hallucinationer (som ved træning dog forsvandt). Lige meget hvilken

løsning, vi brugte, krævede det korrekturlæsning. En mere erfaren programmør end Line ville kunne skrive et OCR-skanningsprogram, som kunne afhjælpe mange af problemerne. Line kunne nok også selv skrive et bedre program, men da projektet både skulle afsluttes i løbet af praktikperioden og havde et afgrænset, men stadig stort, antal spørgelister, blev det på et tidspunkt mere tidssparende at læse korrektur og rette fejl manuelt end at programmere OCR-skanningen bedre. AI var den bedre løsning ift. korrekthed og potentiel hastighed, men det må stadig være en prioriteringssag, om man vil bruge det, da det er et problematisk værktøj, som skader klimaet (jf. f.eks. Strubell et al. 2019) og er bygget på kilder af tvivlsom lovlig herkomst (jf. f.eks. Chang et al. 2023).

Fremtiden?

Digitaliseringen af de næsten 4600 spørgelistespørgsmål er nu færdiggjort og opbevares på Peter Skautrup Centrets server som både excel- og pdf-fil. Vores digitaliseringsprojekt har som udgangspunkt været tænkt som en pilotundersøgelse af mulighederne for en generel digitalisering af centrets materiale. Formålet med projektet har først og fremmest været at undersøge funktionaliteten af de tilgængelige digitaliseringsredskaber som AI og kodning. Begge må siges at have vist gode resultater, hvorfor projektet fortsatte, så alle spørgelistespørgsmål blev digitaliseret. For spørgelistesamlingen åbner digitalisering en mulighed for lettere at søge enkelte spørgsmål frem, hvis man leder efter et specifikt ord, hvor det før var mere besværligt. Det vil også fungere som et overblik over, hvilke spørgsmål det er muligt at få svar på. Enten i spørgelistesvarene eller på jyskordbog.dk. Hvis man ved, hvad der er blevet spurgt om før, kan man nemmere målrette fremtidige dialektundersøgelser, f.eks. en ny indsamling af ord og udtryk på sociale medier. Lignende fordele ville være at finde ved en generel digitalisering af Peter Skautrup Centrets mange samlinger. En digital database ville give flere uden for centret mulighed for at tilgå og bruge dets materialer. Både for fagfolk og den øvrige befolkning. Digitalisering ville desuden muliggøre forskning, hvor det er nødvendigt at kunne behandle store mængder af data samtidigt (jf.

f.eks. Hansen et al. 2024). Digitalisering er desuden vigtigt af hensyn til at fremtidssikre materialerne ved at have kopier flere steder, så alt ikke kan mistes i en ulykke.

Litteratur

- Aarhus Universitet. 2025. Studerende.au.dk. "Generative artificial intelligence (GAI)". Kan tilgås på <https://studerende.au.dk/gai>.
- Chang, Kent K., Mackenzie Cramer, Sandeep Soni & David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. I *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*: 7312-7327. Singapore: Association for Computational Linguistics.
- Det Danske Sprog- og litteraturselskab. 2025. Den Danske Ordbog. "Prompt". Kan tilgås på <https://ordnet.dk/ddo/ordbog?query=prompt&tab=for>.
- Fink, Hans. 2021. "Professionsetik for forskere – forsøg på et overblik". *Tidsskrift for professionsstudier* 33: 6-15.
- Hansen, Inger Schoonderbeek, Mette-Marie Møller Svendsen & Kristoffer Friis Bøegh. 2024. Digitalisering af Jysk Ordbogs seddelsamling. I Louise Holmer, Greta Horn, Hans Landqvist, Pär Nilsson, Eva Nordgren, Emma Sköldberg (red.), *Nordiske Studier i Leksikografi* 16: 99-112. Kan tilgås på <https://tidsskrift.dk/nsil/article/view/143541>.
- Holm, Maja Demetry Thomasen. 2024. Peter Skautrup Centrets spørgelister: Om når jyderne "fistrer, fister, finstrer" omkring. *Ord & Sag* 44: 59-71. Kan tilgås på <https://tidsskrift.dk/ordogsag/article/view/152352>.
- Strubell, Emma, Ananya Ganesh & Andrew McCallum. 2019. *Energy and Policy Considerations for Deep Learning in NLP*. Ithaca: Cornell University Library, arXiv.org.