Automatiserad texttolkning och bearbetning med stora språkmodeller

Källor från Nya Sverige

Gerard Farrell

Denna artikel beskriver en pilotstudie där volymer av material från Riksarkivet rörande kolonin Nya Sverige transkriberades och sammanfattades/översattes med hjälp av automatiserad handskriftstolkning (*Handwritten text recognition*, HTR) och AI-drivna stora språkmodeller (*large language models*, LLMs). Detta genomfördes inom ramen för projektet *Empire: Cromwellian Ireland and the Transformation of the English Atlantic World*, som finansieras av Irish Research Council och leds av professor Micheál Ó Siochrú vid Trinity College i Dublin. *Empire* undersöker Englands (och senare Storbritanniens) växande globala dominans fram till mitten av 1700-talet, med särskilt fokus på Irlands ofta förbisedda roll i denna process, inte bara som en koloni i sig, utan också som en källa till kapital och arbetskraft som bidrog till imperiets tillväxt. En central hypotes i *Empire*-projektet är att den omfattande konfiskeringen av katolskt land på Irland möjliggjorde en makalös kapitalackumulation som i sin tur lade den finansiella grunden för Englands utveckling till ett imperium.

Projektet kombinerar traditionell historisk forskning med en innovativ strategi inom digital humaniora för att analysera de jordägande och finansiella nätverk som uppstod ur denna radikala samhällsomvandling. Lyckligtvis råder det ingen brist på material för att genomföra studien. Mycket av detta material har nyligen återupptäckts och tillgängliggjorts genom projektet *Beyond 2022*, finansierat av Irlands regering och lett av Trinity College i Dublin. Inom ramen för detta projekt skapades Virtual Record Treasury of Ireland (VRTI), en digital rekonstruktion av en stor del av samlingarna från före detta Public Records Office of Ireland som förstördes under irländ-

ska inbördeskriget i juni 1922.[1] Arbetet ledde till upptäckten av en enorm mängd material i form av kopior av förstörda dokument som sammantaget är kända som *the Commonwealth Books,* och som utgör den mest fullständiga bevarade dokumentationen av engelsk kolonial finans och administration under 1600-talet. Tillgången till dessa dokument, varav många inte har studerats på över ett sekel (om någonsin) har skapat både stora möjligheter och utmaningar. Hur ska man på ett effektivt sätt kunna bearbeta och analysera en så överväldigande mängd historiskt material?

En viktig del av projektet är därför att undersöka teknologiska lösningar till denna utmaning, särskilt genom en kombination av HTR-programvara som Transkribus (för att möjliggöra snabb transkription av handskrivna handlingar) och chatbotar som ChatGPT för att bearbeta stora mängder primärkällor. Sådana teknologier gör det möjligt att analysera mönster i en skala som tidigare varit omöjlig. I detta avseende har teknikens tillämpbarhet långt utanför det irländska materialet blivit särskilt tydlig. Ett av huvudmålen med *Empire*-projektet är att utveckla ett jämförande ramverk för att förstå varför vissa imperier lyckades utvidga och konsolidera sin makt medan andra misslyckades. Genom att tillämpa dessa metoder på arkivmaterial från andra koloniala sammanhang strävar vi efter att skapa en bred, datadriven grund för jämförande analys mellan de europeiska imperierna under den tidigmoderna perioden. Som ett första steg i denna riktning genomfördes en pilotstudie där material från Riksarkivet rörande kolonisationsprojektet Nya Sverige transkriberades, översattes och/eller sammanfattades till engelska och samlades i register. Detta fungerar som en utvärdering av läget inom detta snabbutvecklande fält och av teknikens användbarhet för historiska studier, samt som en modell för framtida storskaliga projekt som syftar till att bearbeta arkivmaterial.

Den här artikeln beskriver flera aspekter av projektet. Först ges en kort översikt över den arkivvolym som stått i centrum för projektet, samt hur den passar in i Nya Sverige-samlingen vid Riksarkivet. Därefter beskrivs projektets arbete i mer tekniska termer och de utmaningar som uppstått. Sedan presenteras några exempel från projektet, och avslutningsvis diskuteras metodens potential, både inom vårt eget projekt om tidigmodern engelsk imperialism och i ett bredare europeiskt sammanhang.

1 Peter Crooks & Ciarán Wallace, "The Records of a Country are its Noblest Inheritance: Beyond 2022 – Ireland's Virtual Record Treasury Research Project", *History Ireland* 2020:3; Peter Crooks & Ciarán Wallace, "The Great Untilled Field: The Public Record Office Of Ireland, 1867–1922", *History Ireland* 2021:3; Peter Crooks, Zoë Reid & Ciarán Wallace, "The Virtual Record Treasury of Ireland: A Century of Recovery from the 1922 Four Courts Blaze and Beyond", *History Ireland* 2022:3.

Arkivmaterialet om Nya Sverige

I Riksarkivet finns material om Nya Sverige utspritt över många volymer av Riksregistraturet, samt i form av skrifter i rikskanslern Axel Oxenstiernas samling. I Kammararkivet finns flera volymer med titeln ”Nya Sverige”, liksom material relaterat till de handelskompanier som grundades vid samma tid. De mest sammanhållna samlingarna rörande Nya Sverige återfinns dock i serien ”Miscellanea”, närmare kategoriserad under ”Handel och sjöfart 1500–1800”. Bland underavdelningarna till denna kategori finns nära 200 volymer rörande handelskompanier, där bland annat volymerna ”Söderkompaniet” och ”Skeppskompaniet” innehåller relevant material rörande Nya Sverige. De flesta handlingar av intresse under kategorin ”Handel och sjöfart” finns i en serie benämnd ”Kolonier/Nya Sverige” och omfattar fem volymer. Av dessa volymer rör sig 194b om köp av mark från ursprungsbefolkningen, 195 är en tunn volym om tobakshandeln, 196 innehåller en mångsidig samling rapporter och beskrivningar, suppliker till kronan och brev från guvernör Printz samt omfattande och detaljerade (men svårt skadade) räkenskaper och skuldförteckningar. Volym 197 innehåller brev och kyrkliga handlingar från perioden långt efter att det svenska styret upphört.

Vi beslutade att transkribera och sammanställa volym 194[2] eftersom tidsperioden sammanfaller med den kronologiska ram som är relevant för vårt projekt och då den innehåller det mest varierade och intressanta materialet, både till innehåll och form, med en mångfald av handstilar och språk – främst svenska, men även en betydande mängd nederländska samt några enstaka handlingar på tyska. Det sistnämnda kriteriet var viktigt för att testa Transkribus flexibilitet samt ChatGPT:s förmåga att bearbeta olika typer av text. Volymen innehåller material från 1636–1637 rörande Peter Minuits första expedition och de instruktioner han fick (ff.3r–15v), brev från 1639–1640 med rapporter om logistiska angelägenheter i Göteborg (ff.19r–28v), instruktioner för Johan Printz 1642 samt flera rapporter som skickades till Söderkompaniet (ff.44r–57r), instruktioner till hans efterträdare Johan Risingh och hans dagbok från 1654 (ff.73r–83r), handlingar rörande tvisten 1643–1644 med New Haven-kolonin under George Lambertons ledning, inklusive korrespondens mellan John Winthrop och Johan Printz (ff.158r–169r), dokument om förlusten av skeppet Kattan nära Puerto Rico 1649–1654 (ff.200r–215r) samt diverse petitioner riktade till Axel Oxenstierna 1649–1654 (ff.222r–230r). Volymen avslutas med ett brev från Peter Stuyvesant till Johan Printz, (daterade i Nya Amsterdam den 24 juli 1650) om förlusten av Kattan och politiska händelser i Europa (f.231r).

2 Volym 194 ska inte förväxlas med volym 194b.

Tekniska förutsättningar

I projektet har vi alltså undersökt hur en kombination av Transkribus och ChatGPT kunde effektivisera bearbetning av en manuskriptvolym omfattande 130 folioblad (260 sidor). Förutom på svenska var en betydande del material på nederländska, tyska eller en blandning av båda. Detta är inte förvånande med tanke på hur många nederländare och valloner som var involverade i Nya Sveriges grundande samt det faktum att Delawareområdet utgjorde en mötesplats för kulturer och språk. Det dagliga livet i kolonin innebar att man talade och hörde (om än inte nödvändigtvis förstod) svenska, finska, nederländska och engelska – för att inte tala om diverse irokesiska språk och algonkinspråk. Det är också värt att ha i åtanke det svenska språkets status vid denna tid. Även om Sverige under Gustav II Adolfs ledning hade blivit en militär stormakt, betraktades landet fortfarande som en del av den kulturella periferin. Det fanns en språklig hierarki som i praktiken innebar att en svensk som mötte en nederländsk- eller tysktalande person med stor sannolikhet skulle tala nederländska eller tyska, inte tvärtom. Samma upplevda kulturella rangordning innebar omvänt att en svensk utan tvivel förutsatte att svenska skulle vara kommunikationsspråket i möten med finnar eller dem som talade baltiska språk.[3]

De nederländska och tyska texterna i denna volym väckte en intressant fråga: I hur hög grad kan AI bistå med transkribering och översättning av texter på språk som användaren inte behärskar alls? Författaren (som har svenska som andraspråk) har inte mer än en ytlig förståelse av texter skrivna på nederländska eller tyska, men upptäckte med hjälp av ChatGPT att det ändå gick att åstadkomma en någorlunda användbar transkription av det icke-svenska materialet i volymen. Detta var dock något tidskrävande och väckte viktiga frågor om hur kvalitet ska vägas mot kvantitet i ett sådant projekt. Poängen med att använda AI i detta sammanhang är ju just hastigheten. Om tidsaspekten helt uteslöts ur ekvationen, skulle det knappast finnas någon mening med att använda vare sig Transkribus eller ChatGPT. Åtminstone på den teknologiska nivå vi befinner oss på idag kommer en skicklig mänsklig paleograf eller översättare alltid att kunna utföra arbetet med högre precision. Men precision var inte huvudsyftet med denna övning, och inte heller syftet med dessa verktyg i deras nuvarande form. En transkription av de cirka 260 sidorna i denna volym kan med Transkribus genereras på en timme. Även om resultatet utan tvivel saknar

3 Margareta Revera, ”The Making of a Civilized Nation: Nation-Building, Aristocratic Culture and Social Change”, i *New Sweden in America*, Carol Hoffecker (red.), Newark 1995, s. 32–36; Michael Roberts, *The Swedish Imperial Experience, 1560–1718*, Cambridge 1979, s. 21–22.

den noggrannhet en mänsklig transkriberare skulle kunna uppnå, skulle en människa åtminstone behöva flera veckor för att utföra samma uppgift. I verkligheten, med hänsyn till sömn, arbete och andra åtaganden, skulle det troligen röra sig om flera månaders arbete. Det som är potentiellt avgörande är därför den tidsvinst som uppnås genom att ett första utkast kan skapas så snabbt. Detta första Transkribusutkast bör ses som en grund att bygga vidare på, snarare än som ett färdigt resultat.

Val av rätt Transkribusmodell

Hastighet var alltså det främsta målet i vår transkribering, snarare än noggrannhet. Noggrannhet betraktades som en faktor att optimera inom ramen för Transkribus nuvarande arkitektur, utan att vi själva försökte förbättra denna kapacitet. Projektet har med andra ord inte utvecklat några HTR-modeller specifikt anpassade till detta material (något som utan tvivel skulle förbättra resultaten), utan har använt de offentligt tillgängliga modeller som redan finns i Transkribus. Avgörande för att uppnå så goda resultat som möjligt i transkriberingen var därför valet av den mest lämpliga modellen för varje skrivstil. En modell i Transkribus är ett tränat neuralt nätverk som har lärt sig att koppla visuella mönster i handskriften, så som pixelmönster och pennstreck i de uppladdade bilderna, till sannolika bokstavs- och ordsekvenser baserat på exempel från en viss skrivstil. Den arkitektur som för närvarande oftast används för att träna dessa modeller är PyLaia, utvecklad vid Universitat Autònoma de Barcelona. Den lär sig genom att jämföra exempel på korrekta transkriberingar (så kallad *ground truth*) med sina egna gissningar om vilka mönster som mest sannolikt motsvarar vilka bokstäver och i vilken kontext de förekommer. Varje gissning jämförs med det korrekta svaret, felmarginalen kvantifieras och modellen justeras så att den blir mindre felaktig vid nästa försök. Denna process upprepas tusentals gånger tills modellen når en punkt där den inte längre förbättrar sin noggrannhet vid jämförelse med ett särskilt korrekturunderlag (*validation set*) som inte använts i träningen men har reserverats just för detta syfte.[4]

Att vara noggrann i valet av modell för varje handstil kan på sikt spara avsevärd tid. Den mest effektiva metoden har visat sig vara att först få en överblick över materialet, bedöma hur mycket handstilen varierar, identifiera de olika skrivhänderna och därefter gruppera dessa i breda (men inte alltför breda) kategorier baserade på likhet. I denna volym kunde tre skrivhän-

4 För en mer ingående introduktion till Transkribusapplikationen och dess användare, se Günter Mühlberger m.fl.,"Transforming Scholarship in the Archives Through Handwritten Text Recognition: Transkribus as a Case Study", *Journal of Documentation* 2019:5.

der urskiljas, och dessa var tillräckligt olika för att motivera användning av skilda modeller. Beroende på hur varierat materialet är, är det sällan tidsmässigt motiverat att gå igenom alla Transkribus offentliga modeller – cirka 400 stycken i mars 2026 – i jakt på den perfekta modellen för varje enskilt dokument. I vissa fall kräver en särskilt idiosynkratisk handstil en mer specialiserad modell, om en sådan finns tillgänglig. Som i denna volym måste man även ta hänsyn till vilket språk en text är skriven på, eftersom de flesta Transkribusmodeller är språkberoende. För de nederländska och tyska texterna i denna volym gav inte en PyLaia-modell bäst resultat, utan modellen Text Titan I, en av de nya generationerna transformerbaserade *super models* som tränats av Transkribus på ett betydligt större och mer varierat textunderlag än vad som varit fallet med PyLaia.

Utöver att känna igen visuella mönster och tecken, använder sig transformermodeller av en neutral språkmodellskomponent som aktivt försöker förbättra transkriptionen genom att förutsäga ord baserat på mönster i träningsmaterialet. Dessa modeller har redan visat sig prestera bättre än PyLaia i större skala, med färre sidor med felaktigheter.[5] Även om transformermodeller är mer flexibla och statistiskt träffsäkra, kan viss precision gå förlorad i mötet med texter som på något sätt är idiosynkratiska, antingen vad gäller handstilen eller det språk respektive den terminologi som används och som inte fanns i de texter modellen har tränats på. Den största utmaningen vid användningen av en neuralspråklig komponent på detta sätt är dess benägenhet att ingripa alltför aggressivt vid tolkningen av tvetydiga ord eller fraser, vilket kan resultera i så kallad hallucinerad text. Ordet *Virginien*, till exempel, förekommer ofta i denna volym och skrevs, liksom de flesta person- och ortnamn, ofta med särskild omsorg av skrivaren eftersom dessa ord måste vara tydligt läsbara för samtida läsare. Utan språkmodulen aktiverad kunde Transkribus ofta känna igen ordet tecken för tecken och transkribera det korrekt. Med språkmodulen aktiverad gjordes däremot ett försök att matcha ordet mot modellens interna ordförråd, vilket ledde till att *Virginien* förvreds till bland annat *vimen, Ligiaren, Driginien, invissinen* och *Nirainen*.

Modellen ”Text Titan I” tränades på ett mycket stort textunderlag på tyska, engelska, nederländska, franska, finska och svenska. I praktiken visade det sig dock att resultaten för nederländska och tyska var avsevärt bättre än för svenska. Detta är dock inte förvånande, när den relativa mängden text på svenska (7 procent jämförs med den på tyska (29 procent)

5 READ-COOP, ”Introducing Transkribus Super Models: Get Access to the Text Titan I”, https://blog.transkribus.org/en/introducing-transkribus-super-models-get-access-to-the-text-titan-i, (17/4 2025).

och nederländska (21 procent).[6] För det svenskspråkiga materialet visade sig modellen Swedish Lion I ge de bästa resultaten. Denna modell är en relativt bred allmänmodell baserad på över 15,5 miljoner ord från texter daterade mellan 1600- och 1800-talet.[7] Ett antal texter i denna volym var skrivna med en mer ”gotisk” handstil, och för dessa gav en annan modell, Swedish 17th Century, bättre resultat än Swedish Lion. Denna modell är tränad på häradsrätts- och kyrkböcker från Savolax i mellersta Finland.

Liksom i transformermodeller finns även i PyLaia-modeller möjligheten att använda en språkmodulkomponent (dock icke-neural och mindre avancerad än motsvarigheten i transformermodeller) för att gissa ord utifrån kontext. De bästa resultaten uppnåddes dock med språkmodulen avaktiverad, särskilt vid transkribering av geografiska beskrivningar av Delawareområdet. Som skaparna av Swedish 17th Century-modellen själva har påpekat, begränsar det faktum att modellen är tränad på material från ett relativt snävt geografiskt område dess förmåga att tolka person- och ortnamn, något som bekräftades i denna undersökning. PyLaia-språkmodeller är utan tvekan effektiva när de används för att känna igen samma typ av material som de är tränade på – i detta fall de lokala förhållandena i Savolax under 1600-talet – men vid tillämpning på texter som skiljer sig markant i innehåll tenderar de snarare att försvåra än underlätta arbetet.

Granskning och korrigering av transkriptionerna

Den råa transkriptionen från Transkribus bör ses som ett utkast att utgå ifrån vid framställningen av en transkription som är redo för att kunna tolkas av ChatGPT. En perfekt transkription är inte nödvändigtvis målet, och i många fall vore det heller inte möjligt med tanke på materialets svårighetsgrad och dess dåliga fysiska skick (bleknad text, sönderrivna eller brända sidor). Målet är snarare en transkription som är tillräckligt sammanhängande och fullständig för att ChatGPT ska kunna förstå den. En snabb genomgång och korrigering är därför värd att genomföra inför detta steg. Det finns några manuella åtgärder som har visat sig göra en märkbar kvalitativ skillnad när det gäller att generera en mer sammanhängande parafras eller översättning i ChatGPT. Den första är att dela upp texten i enlighet med de ursprungliga styckeindelningarna. Dessa fångas ofta inte upp av Transkribus, antingen

6 Tack till Andy Stauder från READ-COOP för denna information.

7 Swedish Lion var resultatet av ett samarbetsprojekt mellan riksarkiven i Sverige och Finland, i partnerskap med Stockholms stadsarkiv, Jämtlands läns fornskriftsällskap, medborgarforskare samt forskare från Stockholms och Uppsala universitet. Se READ-COOP, ”Creating the Swedish Lion: Text Recognition Model”, www.transkribus.org/success-story/creating-the-swedish-lion-%E2%85%B0-model (25/4 2025).

för att de helt saknas eller för att de inte är tydligt markerade i originalet. För det andra, eftersom egennamnen (som ofta skrevs extra tydligt för den samtida läsarens skull) ofta ställer till problem för Transkribus, är det värt att gå igenom texten och manuellt korrigera dessa. Detta är inte särskilt tidskrävande, eftersom de är relativt få och ofta sticker ut visuellt i texten.

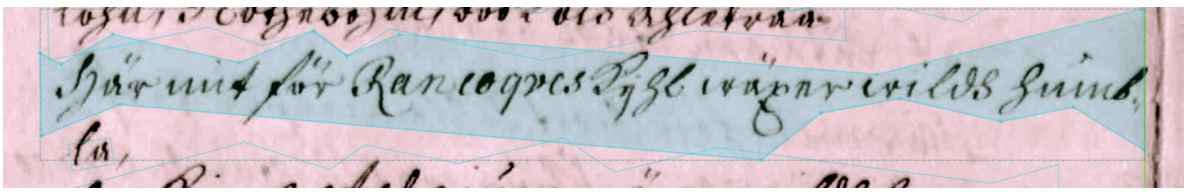
Även om detta projekt i hög grad prioriterade hastighet, lades stor vikt vid att genomgående finna en balans mellan kvalitet och kvantitet. I varje moment ställdes frågorna: Hur noggrant är tillräckligt noggrant för våra syften? Vilken förlust i precision är acceptabel i utbyte mot ökad hastighet och effektivitet? Hur avgör vi när för mycket noggrannhet har offrats för snabbhet, eller tvärtom? En viktig insikt var att snabbhet inte alltid behöver innebära en kompromiss med noggrannheten. I stället kan avvägningar göras på andra områden, till exempel genom att minska detaljnivån snarare än att insistera på ordagrann transkription. I vissa fall har materialet varit så svårtillgängligt att den transkription som producerats av Transkribus i det närmaste varit oanvändbar som underlag för maskinöversättning. Men även i sådana fall är inte allt förlorat. ChatGPT har visat sig vara ett särskilt användbart verktyg för att ändå extrahera en viss innebörd ur i stort sett osammanhängande texter och kunna tillhandahålla någon form av användbar sammanfattning. Vid närmare granskning av dessa utmanande texter framgår det i nästan varje fall att en fullständig transkription skulle vara möjlig att åstadkomma, men endast genom att lägga ned timmar på att tyda varje enskild bokstav och pussla ihop orden med hjälp av till exempel *Svenska Akademiens ordbok*. Det bedömdes dock som mer ändamålsenligt att i dessa fall nöja sig med att sammanfatta innehållet och överlåta en mer detaljerad transkription till den forskare som i framtiden kan ha nytta av denna sammanfattning som ingång. I andra fall, där en text är särskilt tydlig, kan det vara fullt genomförbart att transkribera den ord för ord, och i sådana fall kan det också vara möjligt att tillhandahålla en direkt översättning snarare än en sammanfattning, beroende på materialets karaktär. En av fördelarna med att producera ett register som detta är just att detaljnivån kan anpassas efter transkriptionens kvalitet och dokumentets betydelse.

Bearbetning i ChatGPT

När transkriptionen har förbättrats så mycket som möjligt inom givna tidsramar, kan den bearbetas i ChatGPT. I vårt projekt ingick översättning av dokumenten till engelska som en del av processen, men detta var inte den viktigaste aspekten av detta steg. En översättningstjänst som Google Translate hade räckt om syftet enbart varit att översätta, men LLMs som ChatGPT kan göra betydligt mer än så. I detta fall innebar bearbetningen

att parafrasera texten – i vissa fall för att göra den mer begriplig när transkriptionen var så full av luckor att den gränsade till osammanhängande, i andra fall enbart för att göra texten mer kortfattad. Implicit i denna process låg också en modernisering av texten från 1600-talssvenska som ett förberedande steg inför översättning till engelska. Några tester genomfördes även där texten först översattes till modern svenska och därefter till engelska som två separata moment, i stället för att låta ChatGPT hantera hela konverteringen i ett steg. Även om detta gav olika resultat, kunde ingen tydlig skillnad i kvalitet, vare sig positiv eller negativ, konstateras.

Denna förmåga att utvinna en viss grad av sammanhang ur ofullständiga eller bristfälliga transkriptioner var kanske den mest imponerande aspekten av ChatGPT:s prestanda i detta projekt. Att gissa sig till saknade eller otydliga ord utifrån kontext med hjälp av sannolikhetsbaserade språkmodeller är själva kärnan i denna teknologi. Detta är inte platsen för en djupgående diskussion om hur LLMs fungerar (något som tillverkarna av ChatGPT, OpenAI, inte har gjort helt transparent) men i grova drag bygger processen för ordprediktion på samma principer som i Transkribus språkmoduler, om än tränad på en betydligt större datamängd, enligt uppgift upp till en petabyte.[8] Ett exempel på hur detta fungerat effektivt i vårt projekt är värt att lyfta fram. Följande bild illustrerar detta:



Exempel på handskriven text som transkriberats genom Transkribus.

Transkribus transkriberade denna rad som: ”härmit för Kancoqves Pyhl wäxer wildh huus”. Före bearbetning i ChatGPT förbättrades den manuellt till: ”Här mit för Rancoqves kyhl wäxer wildh huus”. ”Kyhl” är ett mycket vanligt nederländskt ord (*kil*) i texter rörande Nya Sverige och syftar på ett vattendrag eller en vik. ”Rancoqves” är en äldre stavningsvariant av det som idag kallas Rancocas Creek i New Jersey, några kilometer uppströms från Philadelphia. Vid denna punkt kunde meningen på engelska översättas till: ”Wild [houses?] grow by the banks of Rancocas Creek”.

Detta ligger nära en tolkning, men är inte helt tillfredsställande, eftersom ordet ”hus” knappast är rimligt om orden ”wildh” och ”wäxer” är korrekt

8 Seifeur Guizeni, ”Decoding the Enormous Scale of GPT-4: An In-Depth Exploration of the Model’s Size and Abilities”, seifeur.com (23/4 2025).

lästa. Trots att ChatGPT inte hade tillgång till någon bild av originalmanuskriptet lyckades modellen ändå lösa meningen på egen hand. Detta är särskilt imponerande med tanke på att både författaren och layoutanalysen i Transkribus hade missat den följande raden helt, vilken endast innehöll stavelsen ”la”, som skulle ha avslöjat att ”huus” på föregående rad i själva verket var ”humb”, alltså början på ordet ”humbla”, det vill säga humle. ChatGPT drog själv denna slutsats, vilket gjorde det möjligt att låsa upp meningen:

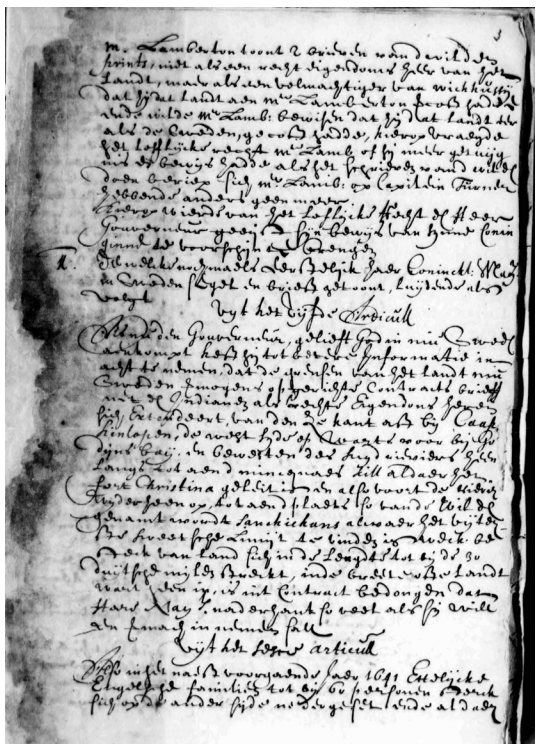
> Här mit för Rancoqves kyhl wäxer wildh humbla
> Här mitt framför Rancocas kil växer vild humle
> Wild hops grow by the banks of Rancocas Creek

Det bör noteras att förmågan att förutsäga innehåll utifrån kontext, som gör LLMs så användbara, också kan orsaka problem. En alltför ”invasiv” nivå av tolkning kan som sagt leda till att text uppfinns ur tomma intet, alltså det så kallade hallucinationsfenomenet. Utan tillsyn har ChatGPT visat sig kapabel att fylla i luckor i texten i förvånansvärt hög grad. Även om detta ytligt sett kan framstå som någon form av kreativitet, är det i själva verket ett resultat av modellens beroende av statistiska mönster snarare än faktisk förståelse. Eftersom LLMs inte ”vet” saker i egentlig mening, utan enbart förutsäger sannolika sekvenser av ord, kan de generera information som låter trovärdig men som i själva verket är helt påhittad. Att *låta* trovärdig är, i viss mån, själva kärnan i vad modellen gör. Även om utvecklare arbetar för att minska detta problem på programmeringsnivå, kan en historiker som använder ChatGPT som arbetsverktyg under tiden mildra risken genom att reglera graden av spekulation med hjälp av noggrant formulerade instruktioner (prompter). Snarare än att helt förbjuda spekulation har det visat sig mer ändamålsenligt att instruera modellen att förklara sitt resonemang i de fall där tolkningen uppenbart är mer spekulativ. Dessa motiveringar kan sedan användas som stöd vid granskningen av texter, genom att jämföra med originaltexten för att avgöra om innebörden återgetts korrekt.

Exempel: Rättsprocessen vid Fort Christina

Det återstår att presentera ett exempel från registret som producerades av detta projekt. Följande sida är hämtad ur protokollen från en rättsprocess vid Fort Christina den 10 juli 1643 mellan guvernör Johan Printz och George Lamberton, en engelsk handelsman från New Haven. Bakgrunden till rättsprocessen var närvaron av nybyggare från New Haven under Lambertons ledning. Nybyggarna hade anlänt till området 1641 i syfte att få tillgång till pälshandeln. De hade upprättat bosättningar nära dagens Salem och vid

floden Schuylkill, och hävdade att de köpt marken från Lenaperna. Deras upprepade ”försäljning” av samma mark till alla tre europeiska makter som tävlade om inflytande längs floden var troligen mindre ett försök till bedrägeri än en följd av missförstånd, grundat på européernas uppfattning om markägande som något som innebar exklusiva rättigheter – en föreställning som ursprungsbefolkningen inte delade. Den engelska närvaron väckte naturligtvis oro både bland svenskarna och nederländarna, och de senare avlägsnade bosättningen vid Schuylkill med våld 1642. Eftersom Sverige och England vid denna tid var i fred, lät guvernör Printz den engelska bosättningen längre nedströms vara för tillfället. När en möjlighet att gripa Lamberton, som fortsatt handla längs floden, visade sig i juni 1643 (då ett rykte spreds om att han planerade att uppvigla Lenaperna till att massakrera både svenskarna och nederländarna) tillfångatog Printz honom och ställde honom inför rätta.[9] Det är denna rättegång, där Lamberton formellt åtalades för den påstådda sammansvärjningen men där fokus snarare låg på att bestrida engelsmännens anspråk på att få bosätta sig vid floden, som handlingen ifråga dokumenterar.



RA 757/16/C/3/194, f.162r.

9 Johnson, *Swedish Settlements*, volym 1, s. 382–290; Weslager, *New Sweden on the Delaware*, s. 63–66.

Transkribus transkriberade texten, som är skriven på nederländska, på följande vis – rad för rad:

> M. Lamberton toont 2 brieven van d wilders
> Prints, niet als een recht eigendoms heer van het
> Landt, maer als een volmachtiger van wichhüssig
> dat hij dat landt aen mr Lamberton vercoff hadde
> ende wilde mr Lamb: bewisen dat hij dat landt ter
> als de Sweden, gecoft hadde, hierop vraeyde
> het lefflijcke recht mr Lamb of hij meer getuijg
> nis en bewijs hadde als het schreeven van d wilde
> doen beriep sich mr. Lamb: op Capitein Turner
> hebbende anders geen meer
> hierop wierde van het leflijcke recht den heer
> gouverneur geeijst sijn bewijs van huwe Conin
> ginne te voorschyn te brengen.
> De welcke nochmaels eerstelijck haer Coninckt: Mag
> in Sweden segel en brieff getoont, luijdende als volgt
>
> Uyt het vijfde Articull
>
> Als nu den Gouverneur, gelieft God in niu Swede
> aenkompt heeft hij tot betere informatie in
> acht te nemen, dat de grensen van het landt niu
> Sweden gemogens opgerichte Contracts brieff
> met den Indianen als rechte Eigendons heren
> sich extendeert van den ze kant aff bij Caap,
> Hinlopen, de west sijde op wazt voor bij Go
> dijns bay, en bewesten des suyd riviers heen
> langs tot aen d' Minquaes Kill aldaer het
> Fort Christina geleit is en also voort de riveren
> wyderheen op, tot aen d plaets so van de wilde
> genamt wordt Sanchickans alwaer het vijter
> ste Sweetsche limuyt te vinden is welck be
> steck van land sich in de lengdte tot bij de 30
> duijtsche mijlen streckt, inde breete ofte landt
> wart heen in, is int Contract bedongen dat
> Hare May, naderhant so veel als sij will
> de smach in nemen sall.
>
> Uyt het sesce articull.
>
> Also in het naest voorgaende jaer 1641 Ettelijcke
> Engelsche famities tot bij 60 personen sterck
> sich op da ander sijde nedergeset ende aldaer

Översatt till svenska får sidan följande innehåll – nu utan indelningar i rader:

> Herr Lamberton framlade två brev från vildarnna, inte såsom från de rättmätiga ägarna av landet, utan som från en auktoriserad representant som påstod han hade sålt landet till honom. Herr Lamberton önskade bevisa att han hade förvärvat landet vid samma tidpunkt som svenskarna hade köpt det. Till detta frågade den ärevördiga rätten herr Lamberton om han hade några ytterligare vittnen eller bevis utöver de skriftliga handlingarna från vildarna. Då åberopade herr Lamberton kapten Turner, eftersom han inte hade några andra. Därefter krävde den ärevördiga rätten att guvernören skulle framlägga sitt bevis från Hennes Majestät drottningen. Denne framlade då återigen, allra först, Hennes Kungliga Majestäts sigill och brev från Sverige, vilka lydde som följer:
>
> Ur den femte artikeln:
>
> Eftersom guvernören – Gud vare tack – har anlänt till Nya Sverige, skall han, för sin bättre orientering, lägga märke till att gränserna för landet Nya Sverige, i kraft av det kontraktsbrev som har upprättats med indianerna såsom de rätta äganderättshavarna, sträcker sig från havskusten vid Kap Henlopen, längs västra sidan vid Godins bukt och längs Södra floden upp till Minquas Kill, där Fort Christina är beläget. Och vidare längs floderna uppströms till den plats som av de infödda kallas Sanhickan, där det tjugofemte svenska gränsmärket finns. Detta jordstycke sträcker sig i längd upp till 30 tyska mil; vad gäller bredden, eller utsträckningen inåt landet, är det i kontraktet fastställt att Hennes Majestät därefter må ta så mycket i besittning som hon önskar.
>
> Ur den sjätte artikeln:
>
> Också under det närmast föregående året, 1641, bosatte sig flera engelska familjer – upp till 60 personer – på motsatta sidan, och där...

En detalj av intresse är av mer indirekt betydelse för den aktuella tvisten. Vid sin ankomst 1638 genomförde svenskarna (under ledning av vallonen Peter Minuit) en ceremoni som de själva uppfattade som ett ”köp” av mark från Lenaperna. Något köpebrev för denna transaktion finns inte bevarat, men detta dokument kan användas för att bekräfta omfattningen av markområdet, både längs flodstranden och inåt land. Den sträcka som här beskrivs längs Delawarebukten (av holländarna kallad Godins Bay, efter Samuel Godin, en av direktörerna i Västindiska kompaniet), förbi Minquas Kill/Fort Christina (dagens Wilmington) och så långt som till Sanhickan (dagens Trenton, New Jersey), motsvarar cirka 30 tyska mil (cirka 200 kilometer). Dokumentet kan även användas för att styrka den inåt landet sträckande omfattningen av svenskarnas ”köp”, som ett komplement till den

något förvirrande formuleringen som användes av vittnen till händelsen, nedtecknad kort därefter, där det påstods att så mycket mark hade ”köpts” som kunde färdas på en dagsresa.[10] Det här föreliggande dokumentet tycks bekräfta att svenskarna tolkade transaktionen som att den omfattade så mycket land inåt landet som de själva ville göra anspråk på – ett hypotetiskt anspråk utan praktiska konsekvenser, men utan tvivel något som skulle ha förvånat Lenaperna som påstods ha gått med på det.

Sammanfattning och utblick

Exemplet ger endast en liten inblick i vad detta material har att erbjuda forskare inom tidigmodern imperiehistoria: topografiska, kulturella och biografiska detaljer, av vilka de mest värdefulla för historikern ofta är just de som är oväntade i förhållande till dokumentens huvudsakliga syfte. Syftet med att undersöka detta material inom ramen för *Empire*-projektet vid Trinity College Dublin var att bygga upp ett jämförande ramverk för att kunna betrakta det engelska och senare brittiska imperiets framväxt i relation till dess konkurrenter, även relativt små aktörer som svenskarna. Även om just denna volym inte gav något större utbyte i det avseendet, var syftet också att utveckla en metodologi. Denna har enligt vår bedömning visat att det är fullt genomförbart att på ett effektivt sätt producera register över stora mängder källmaterial med hjälp av HTR- och LLM-teknik. Det gör det möjligt för oss att bygga detta jämförande ramverk utifrån primärkällor snarare än att förlita oss alltför mycket på sekundärlitteratur, som i stor utsträckning har förbisett den avgörande roll som Irland, och exploateringen av dess mark och resurser, spelade för att ge det engelska imperiet ett avgörande försprång i kapplöpningen om den globala makten.

Nu överväger vi olika möjligheter att använda den metodologi som utvecklats här. En idé har varit att ta fram en samling primärkällor (på engelska kallad *calendar*) som sammanfattar ett brett urval av material liknande det som har analyserats här, men som ännu inte har gjorts tillgängligt. En sådan samling skulle kunna vara flerspråkig och omfatta material på nederländska, franska, tyska och andra språk. Detta väcker viktiga frågor om användningen av maskinöversättning och AI i allmänhet. Ett sådant projekt skulle sannolikt bara vara meningsfullt om det genomförs i stor skala, vilket i sig innebär särskilda utmaningar. Strategier för att hantera dessa måste sannolikt diskuteras, till exempel möjligheten att använda

10 (Tyska) ”alle das land so viel tag-reiß an allen ordern”, vittnesmål av Michell Simons van Sardam och Kalmar Nyckels besättning rörande grundandet av Nya Sverige, Amsterdam 29/12 1638, i Kolonier, Nya Sverige I, Kammarkollegiet Ämnessamlingar, RA 522/09/42, ff.418–419.

crowdsourcad korrekturläsning – något som i sin tur medför ytterligare svårigheter (bristfälligt källmaterial, svårigheter att upprätthålla vetenskaplig standard och så vidare). En annan central fråga är att avgöra vilka källor som ska inkluderas: Vad ska egentligen räknas som relevant för den ”atlantiska världen”?

Projektet har slutligen gett oss en god förståelse för såväl begränsningarna som potentialen hos AI som verktyg i framställningen av historiska data, givet den nuvarande teknologiska nivån. Det har även tydliggjort hur snabbt fältet utvecklas, och hur provisorisk varje bedömning av dess kapacitet nödvändigtvis måste vara. De verktyg som idag står till förfogande är kraftfulla, men långt ifrån ofelbara. AI kan feltolka, ”hallucinera” eller missa historiska nyanser, och även de bästa resultaten kräver noggrann mänsklig granskning. Resultatens kvalitet är fortfarande starkt beroende av kvaliteten på indata och operatörens skicklighet, ett beroende som dock sannolikt kommer att minska under de kommande åren. Trots dessa begränsningar, och även om försiktighet bör iakttas när det gäller att fästa alltför stor tilltro till ChatGPT:s resultat, finns det också en fara i att misstro tekniken alltför mycket. AI är fortfarande långt ifrån att kunna utföra det historiska arbetet åt oss (och det är tveksamt om detta ens vore önskvärt), men det kan användas som vilket annat arbetsbesparande hjälpmedel som helst. Med skicklig användning och noggrann korrekturläsning av de genererade texterna kan det spara mycket tid och arbete. Ibland kan det till och med påvisa sådant som den mänskliga transkriberaren kan missa, särskilt vid bearbetning av stora textmängder under tidspress. Detta frigör mer tid för historikern att läsa och analysera de historiska källorna. En viss grad av skepsis är visserligen sund, men den som gör misstaget att misstro denna teknik alltför mycket riskerar att gå miste om dess betydande fördelar.

Summary

Handwritten Text Recognition and Processing with Large Language Models: Sources from New Sweden

This article describes a pilot study in which records from the Swedish National Archives concerning the colony of New Sweden (1638–1655) were transcribed, translated, and summarised using Handwritten Text Recognition (HTR) via Transkribus and the Large Language Model (LLM) ChatGPT. The project was carried out within the framework of ”Empire”, a research project at Trinity College Dublin that examines Ireland’s role in the building of the English empire during the seventeenth and eighteenth centuries. The aim was to test how effectively these tools could be used to process

this extensive material (over 130 folio pages) with documents in Swedish, German, Dutch, and Latin. The article explains the choice of HTR models, particularly how a combination of Transkribus and transformer-based models such as "Text Titan I" and "Swedish Lion I" were used to optimise speed and reach a level of acceptable accuracy. An important aspect was to assess the ability of LLMs to summarise and translate fragmentary and partially illegible texts, as well as to weigh accuracy against efficiency. The article also presents an example from a 1643 court case between Governor Johan Printz and the English merchant George Lamberton, demonstrating the method's potential for interpreting complex sources. Finally, the author discusses the future potential of this technology and possible avenues for further development, emphasising that although the technology requires human oversight, it is a powerful labour-saving tool that can transform how historical material is made accessible and analysed.

Keywords: Handwritten Text Recognition, Large Language Model, Colonialism, New Sweden, Empire, ChatGPT