# Arkivpodden

## Transkribering av AI och arkiven, avsnitt 2

**Medverkande:** Erik Lenas och Andrëa Grängsjö

**Produktion och musik:** Copyfabriken. Musiken kommer från DayFox via Pixabay och är ”royalty free”.

[Musik]

**Erik Lenas:** Vi vill ju ligga i framkant här. Vi vill utforska de här möjligheterna. Interaktionen mer än historiska texten. AI tillämpad på historisk text.

Och vi tror att vi kommer jobba mycket med det under åren som kommer. Och jag tror att det kommer hända häftiga saker.

**Andrëa Grängsjö:** Du lyssnar på Arkivpodden från Riksarkivet. Jag heter Andrëa Grängsjö och med mig i studion har jag Erik Lenas som är Lead Data Scientist på myndigheten.

Vi ska prata om AI och arkiven. Närmare bestämt mer om vårt arbete med handskriftsigenkänning, datadriven forskning och varför historiska språkmodeller är viktiga.

Men Erik, hur fungerar handskriftsigenkänning, HTR, Handwritten Text Recognition?

**Erik:** Ja, det är en ganska komplicerad process faktiskt. Jag tänkte passa på att ge en liten kort historia.

**Andrëa:** Spännande.

**Erik:** Först fanns OCR, det vill säga Optical Character Recognition, det vill säga att känna igen tryckt text.

Jag ska börja med förresten att säga varför man gör HTR överhuvudtaget. När man har scannat en bild, då har man ju bilden digitalt. Men man har ingen aning om vad som egentligen står på texten, eller på bilden.

För att få reda på det så måste man läsa texten på bilden. Och det är det som HTR och OCR gör. De läser texten på bilden och omvandlar det till digital text. Vilket gör att man kan söka mot texten.

Och om det är en handskrift, till exempel en komplicerad handskrift från 1500-talet, så behöver du inte kunna läsa 1500-talshandskrift, utan du kan få texten digitalt bredvid.

**Andrëa:** Praktiskt.

**Erik:** Och du kan söka gentemot texten.

Hur fungerar då HTR?

Förr i tiden, när det bara fanns OCR i princip, så segmenterade man först upp bilden bokstav för bokstav faktiskt. Så man gjorde små bildklipp, bokstav för bokstav. Och sen tränade man en modell, antingen en AI-baserad modell eller en regelbaserad modell, på att känna igen vilken bokstav det var på den här lilla, lilla bilden. Och sen kopplade man ihop alla de här bokstäverna till en sammanhängande text.

Det här är väldigt svårt med handskrift. För det är väldigt svårt att segmentera handskrift bokstav för bokstav. Och med tanke på hur stor variation det finns i handskriften också, så är det väldigt svårt att avgöra vilken bokstav det är på bilden, när man inte har någon som helst kontext.

Det stora genombrottet för HTR, alltså för Handwritten Text Recognition, kom 2016. När man tränade modeller på att känna igen, alltså AI-modeller på att känna igen, inte en bokstav i taget, utan en hel textrad i taget. Och det man får då, det är kontext.

Men det betyder också att innan man gör den här textigenkänningen, så behöver man segmentera upp bilden i textrader. Det gör man också med hjälp av AI-modeller. Så det är alltså inte bara textigenkänning i Handwritten Text Recognition, utan segmenteringen är minst lika viktig.

Och att segmentera ut textrader i en handskriven, tät 1500-tals text kan vara mer utmanande än man tror. Så segmenteringen är minst lika viktig.

Men åter till själva textigenkänningen nu. När man har en segmenterad textrad, och ska avgöra vad som står på den här textraden, och man tar hela textraden på en gång, då har man kontext.

Så har segmenteringen till exempel missat en prick över i:et, så kan den här AI-modellen ändå sluta sig till att det är ett i, och inte till exempel ett stort I. Och den kan göra det baserat på kontexten, för

den vet vilket ord som det här strecket är en del av, och kan sluta sig till att det är ett i.

Så man kan säga att det stora genombrottet kom 2016, och sedan dess har tekniken och forskningen utvecklats mot vad som kallas transformer-modeller, som till exempel ChatGPT bygger på. Och de är ännu bättre på att hantera kontext.

Ja, det var lite kort om varför man gör handskriftsigenkänning, och lite kort om tekniken.

**Andrëa:** Ja, jättespännande. Men hur länge har Riksarkivet jobbat med HTR då?

**Erik:** Vi har jobbat med HTR sedan 2019, tror jag att det var.

**Andrëa:** Vi var ganska snabbt på då.

**Erik:** Ja, vi var ganska snabbt på, det får man säga. Och det var dels ett projekt som arbetade med poliskammarmaterial från 1800-talet i Göteborg. Och där jobbade man med crowdsourcing för att översätta den här handskrivna texten till digital text manuellt genom experter som transkriberade texten. Och sedan tränade man AI-modeller på den här transkriberade texten, för att sedan AI-tolka andra sjok av texten som man sedan manuellt gick in och rättade genom crowdsourcing.

Resultatet blev en väldigt bra transkription av ett ganska stort material och ett väldigt bra tränings- set för att vidareträna AI-modeller för handskriftsigenkänning.

Parallellt med det här crowdsourcing-projektet, som det var, så genomförde vi ett internt Proof of Concept-projekt på Riksarkivet som kallades för AIRA II, där vi undersökte möjligheten att automatiskt extrahera information ur scannade personakter. Och det här var ett pilotprojekt för den typen av indexeringsprojekt som jag pratade om i förra avsnittet.

Och vi visade här att det var fullt möjligt att med hjälp av segmenteringsmodeller och handskriftsigenkänningsmodeller utvinna information ur scannade arkivhandlingar, formulär så att säga. Vilket banade vägen för senare indexeringsprojekt som vi kallar det, alltså effektiviseringsprojekt där vi förkortar handledningstiden för våra ärenden från allmänheten.

**Andrëa:** Coolt! Absolut.

Man måste ha lagt mycket tid på det här att transkribera och att kontrollera det som var tränat sen. Alltså de texterna. Det måste ha gått åt en hel del tid för många personer.

**Erik:** Ja, precis. Och det är det som är så fantastiskt med crowdsourcing-projekt. Att det finns personer som gör det här och som är experter på det här och som gör det så bra. Och som skapar den här träningsdatan så att vi ska kunna…, dels för att vi ska kunna utföra vårt myndighetsuppdrag men också för att vi ska kunna använda den här datan för att tillgängliggöra material för forskning på ett annat sätt. Så det är ett fantastiskt samarbete. Och det är en fantastisk projektform, de här crowdsourcing-projekten.

**Andrëa:** Härligt. Men är det svårare med äldre handskrift än en modern handskrift?

**Erik:** I princip inte. Kanske lite kontraintuitivt mot vad man tror. Men allt handlar om träningsdata. Det kan till och med vara så att det nästan är enklare med äldre handskrift för den är mycket prydligare. Har man bara träningsdata så funkar det bra att träna från vilken epok som helst.

Men det är klart att det är svårare att ta fram träningsdata för medeltida latin. För det är färre som kan läsa det och färre som kan producera den träningsdatan. Så i princip inte. Men det gäller att ha träningsdata. Och det finns ju även färre texter ju längre tillbaka i tiden vi går.

**Andrëa:** Det är viktigt att ta vara på de som kan det så att vi kan använda det framåt. Det är ju toppen. Men jag tänker på hur tillgänglig gör vi den HTR-ade texten? Hur får vi ut det?

**Erik:** Just nu i år, 2024, så har vi ett projekt på Riksarkivet som heter HTR-publicering. Det går under det interna namnet i alla fall. Det handlar om att utveckla vår bildsöktjänst med digital text. Alltså med HTR-ad text. Och utveckla en sökfunktionalitet så att man kan skriva en fritextsök. Få upp resultatet och få resultatet markerat på bilden också. Vilket erbjuder helt nya möjligheter. Och det här är ju en funktionalitet som vi hoppas på att kunna utveckla framöver mer och mer. Och tillämpa mer AI-baserad sökfunktionalitet på det här också.

Man ska tänka att när man söker mot en 1500-talstext till exempel så är söksträngen på modern svenska, men texten är på 1500-talssvenska. Vilket innebär ett problem. Speciellt om man bara har

vad man kallar för keyword search. Det vill säga att man matchar ord med exakt samma stavning. För stavningen skiljer sig ganska mycket från 1500-tal och nutid. Det här finns det lösningar på, men de är AI-baserade.

**Andrëa:** Åh, vad spännande. Jag vill veta mera. Jag tänker på, i det här arbetet. Vilka arkiv kommer vi att börja med nu?

**Erik:** Vår plan under 2024 är att köra Svea Hovrätt. Vi kommer att köra ungefär 500 000 sidor, tror jag, löptext Svea Hovrätt. Och det kommer att bli vårt pilotprojekt på uppskalad HTR, kan man säga. Vi skriver just nu den kodbas som vi kommer att använda.

Som vi gör nu, så kontinuerligt när vi får tillgång till mer träningsdata så tränar vi bättre modeller. Men vi kommer att träna modeller specialiserade på just Svea Hovrätt. Och så kommer vi att köra de modellerna i vår kodbas, uppskalat. På våra egna GPUer. Förlåt, nu använder jag en sån där term igen. Som jag bara som data scientist förutsätter att alla kan.

**Andrëa:** Vad är en GPU?

**Erik:** GPU betyder Graphical Processing Unit. Och det skiljer sig från en CPU som finns i alla våra laptops. Den skiljer sig för att den är väldigt bra på att parallellisera vektormultiplikation. Alltså göra många vektormultiplikationer samtidigt. Och alla AI-beräkningar är vektormultiplikationer.

En CPU måste göra dem sekventiellt. Det vill säga en efter en efter en. Men en GPU kan göra massor samtidigt. Alltså så är GPU:er mycket snabbare på AI-beräkningar än vad en CPU är.

Ni kanske har hört Nvidia som gjorde GPU:er åt alla gamare här i världen. De har blivit väldigt stora och väldigt rika för att de gör GPU:er. Alla företag, myndigheter och institutioner som jobbar med AI behöver GPU:er.

**Andrëa:** Härligt, jag får lära mig saker hela tiden. Tack för att du förklarade det.

**Erik:** Ingen fara.

**Andrëa:** Vad är nästa steg efter HTR?

**Erik:** Det var lite det jag kom in på när jag pratade om AI-baserad sökfunktionalitet. För att med hjälp av HTR omvandla texten från bild till digital text är bara första steget.

Vi vill skapa en interaktion med texten. Det vill säga att man ska kunna skriva en forskningsorienterad fråga i fritext. Och få svar direkt från det här 1500-talsmaterialet. Men för att man ska göra det så är det ett antal saker man behöver. Man behöver kunna korrelera en söksträng eller sökfråga på nutidssvenska med 1500-talstext.

Då behöver man språkmodeller, alltså typ ChatGPT liknande modeller som också är tränade på historiskt text. För att antingen översätta 1500-talstexten till modern text. Och sedan söka mot den moderna texten men få resultaten i den äldre texten. Eller att språkmodellen klarar av att göra det här åt en. Via att den är tränad på modern text och historisk text.

Hur som helst så är det här ett ganska outforskat fält. Och vi vill ju ligga i framkant här. Vi vill utforska de här möjligheterna. Interaktionen med den historiska texten, AI tillämpad på historisk text. Och vi tror att vi kommer jobba mycket med det under åren som kommer. Och jag tror att det kommer hända häftiga saker.

**Andrëa:** Det tror jag med, det låter så. Absolut. Men jag tänkte, bara för sakens skull höll jag på att säga. Vad är en språkmodell egentligen?

**Erik:** En språkmodell är en modell som förutsäger nästa ord baserat på de orden som har kommit innan. En språkmodell är inget nytt. Det har funnits länge. Man behöver inte alls använda AI för att bygga en språkmodell. Man kan använda rena statistiska beräkningar. Men det som är nytt med de här nya språkmodellerna, det är att de kan tränas på, som vi kallar det, oannoterad data.

Det vill säga att man kan ta all text på internet och så kan man ge de här språkmodellerna en mening. Maska ut ett ord i den här meningen. Alltså dölja ett ord. Och sen säga att nu ska du förutsäga vad det här dolda ordet är för någonting. Och så vet vi ju svaret. Så har den fel, då säger vi till den. Du hade så här mycket fel och då justerar den någon av sina miljarder parametrar för att ha lite mer rätt nästa gång. Och så upprepar man den här processen med all text som finns på hela internet.

**Andrëa:** Och det är mycket!

**Erik:** Det vill säga att det är en träning där man behöver kanske 10 000 GPU:er som kör i ett halvår. Men när den har gjort den här träningen, plötsligt så kan den göra alla de här konstiga sakerna som ChatGPT kan göra. Nämligen förstå språk på ett sätt som man inte riktigt begriper. Men det enda den gör, det är att förutsäga nästa ord utifrån de ord som har kommit innan.

En väldigt simpel träningsuppgift som man definierar åt den. Men bara genom att lära sig den uppgiften bra så kan den göra väldigt mycket. Inte allt naturligtvis. Den har många begränsningar fortfarande. Men förvånande mycket ändå skulle jag säga att ChatGPT kan göra utifrån den väldigt simpla uppgiften den får.

**Andrëa:** Jättekul, jätteroligt. Jag tänker på det här med språkmodeller och historiska språkmodeller. Varför är de viktiga? Det har du ju pratat om.

**Erik:** Jo, jag har ju pratat om det lite redan. De är viktiga för att en språkmodell inte bara ska förstå modern svenska, utan även historisk svenska.

En språkmodell tränad på modern svenska kan ju förstå mycket om den kontext som finns idag. Och de sammanhang som finns idag. Men med hjälp av en historisk språkmodell kan vi analysera historiska texter på ett helt annat sätt. Och den kan även förstå språklig utveckling över tid.

Förstå är ett för starkt ord där, skulle jag säga. Jag skulle inte säga förstå, men med hjälp av den språkmodellen kan en forskare bättre förstå språklig utveckling över tid.

**Andrëa:** Ja, det är fantastiskt vad som kan komma ut. Men jag tänker, när är vi klara med det här jobbet på Riksarkivet, tror du?

**Erik:** Vi är aldrig klara. Jobbet med att tillgängliggöra våra arkiv och utvinna information ur våra arkiv och bevara våra arkiv, det är jobbet ju aldrig klart. Utan i takt med att nya horisonter uppstår och nya möjligheter uppstår så uppstår även nya möjligheter för Riksarkivet. Och arbeta vidare med våra definierade uppgifter. Så att jobbet blir aldrig klart.

Och vi kommer aldrig till någon slutpunkt där vi kan säga att nu är vi färdiga med att tillgängliggöra våra arkiv. Utan det här är ju en kontinuerlig process där våra användare, såväl som vi, är högst involverade. Det är en dialog med vår historia och den tar ju aldrig slut.

**Andrëa:** Och en stor del i det är ju också samarbetet med forskningen och vad som driver forskningen. Vi pratade lite kort om datadriven forskning. Kanske kan du utveckla lite grann där?

**Erik:** Ja, låt mig börja med att säga att det centrala i all humanioraforskning, humanistisk forskning, är en intressant frågeställning. Det kommer man inte ifrån. Oavsett om man närstuderar texter eller om man bedriver datadriven forskning så behöver man en intressant frågeställning.

Men datadriven forskning innebär att man har tillgång till en helt annan mängd text på en gång. Och man kan ställa andra typer av frågor mot texten eftersom man har en sökingång mot texten. Som jag har sagt förut, det här kommer aldrig ersätta närstudiet av texter. Såklart inte. Det kommer alltid vara superviktigt.

Men det erbjuder en annan typ av frågor att ställa mot texter. Ekonomisk forskning, statistisk forskning, forskning hur globala kontexter har utvecklats över tid. Vi kommer inte bara ha tillgång till våra svenska historiska texter, utan samma process som sker på Riksarkivet sker på kulturminnesinstitutioner över hela världen just nu. Och de språkmodeller som utvecklas just nu förstår nästan alla världens språk.

Förlåt, det gör de inte. De förstår inte alla världens språk. Det finns väldigt många språk de inte förstår. Men de kommer utvecklas vidare. Man kommer kunna korrelera texter från hela världen och förstå samband på ett mycket större plan. Och analysera de sambanden. Det är väl det som är datadriven forskning. Att man har tillgång till en sån mängd data.

Men att ha en intressant frågeställning är minst lika viktig. Man ska inte bara flasha ny teknik och sen svara på en ointressant fråga. Det är inte kul forskning. Utan allt hänger fortfarande på frågan vi ställer till vår historia.

**Andrëa:** Där kan vi få ut massor med svar i framtiden. Just det här som du pratar om, det globala. Att kunna komma åt, att kunna köra mot material i olika länder. Och att koppla ihop det och att se samband eller skillnader. Det är ju oerhört spännande.

**Erik:** Verkligen.

**Andrëa:** Men det här med arbetet med att tillgängliggöra arkiv tar ju aldrig slut. Jag måste säga att jag tycker att det känns ganska

betryggande. Med tanke på att vi båda jobbar på Riksarkivet och vi vet hur mycket material det finns. Och vi vet hur mycket material som vi får in allt eftersom.

I vårt land så har vi ju turen att vi har en lång historia och en lång tradition av att spara, särskilt officiella, handlingar. Och det känns jättespännande om man tittar framåt. Det kommer bara att bli mer och man kommer att kunna ha användning för det i framtiden och man kan se väldigt långt bakåt. Så det är dåtid, nutid, framtid. Det är ju jättespännande.

**Erik:** Absolut.

**Andrëa:** Vill du säga någonting mer om just det här med tillgängliggörandet? Eller ska vi knyta ihop?

**Erik:** Jag kan bara förstärka att det aldrig tar slut. Såklart. Vår process med att förstå oss själva, vår samtid, vår historia och kunna blicka framåt mot vår framtid. Det är en process som aldrig tar slut. Och tur är ju det.

**Andrëa:** Ja, härligt. Men tack Erik. Vi fortsätter då arbetet med att tillgängliggöra arkiven *forever*.

**Erik:** Ja, så är det.

**Andrëa:** Tack för att du har lyssnat på Arkivpodden. Podden som vill sprida kunskap, tankar, kultur och öka intresset för arkiv och samhällsviktig information.

Vill du veta mer om Riksarkivet? Gå in på riksarkivet.se.

Och vi som har varit med i det här avsnittet är Andrëa Grängsjö och Erik Lenas från Riksarkivet.

[Musik]