AI FÅR PIPPI PÅ ASTRIDS KRÅKFÖTTER

ASTRID LINDGRENS "KRUMELUNSAR"

Att lära datorer läsa handskriven text är ett växande forskningsområde. Med artificiell intelligens tar sig svenska forskare an både klartext och stenografi, som i Astrid Lindgrens manus.

AI FÅR PIPPI PÅ ASTRIDS KRÅKFÖTTER

Av MATS KARLSSON
Foto EVA DALIN

Att kunna söka fritt i svårlästa gamla handskrifter är en önskedröm för släktforskare och historiker. Med artificiell intelligens, AI, kan det bli möjligt. Men vägen dit är lång. Den går bland annat via Astrid Lindgrens manuskript.

Astrid Lindgren lämnade efter sig 670 anteckningsböcker med manuskript till sina böcker om Pippi Långstrump, Emil i Lönneberga och hennes andra karaktärer. Att försöka följa hennes kreativa processer genom att studera hur hon strök och ändrade i sina manus är en diger uppgift, inte minst för att hon stenograferade på ett mycket svårtytt sätt.

– Det skulle krävas många stenografer under väldigt lång tid för att skriva av manusen i klartext. Därför försöker vi lösa det med artificiell intelligens. Såvitt vi vet har ingen använt AI på ett organiserat sätt för att tolka stenografi, säger Malin Nauwerck, forskare i litteraturvetenskap vid Svenska barnboksinstitutet, som förvarar några av manuskripten.

Hon jobbar nu med *Bröderna Lejonhjärta* med hjälp av volontärer som kan stenografi. De läser inskannade sidor ur manuskriptet, gör rutor kring enskilda ord och anger vad det står i klartext. Sidan läses in i en dator igen, som kopplar ordbilderna i rutorna till ordet i klartext.

Malin Nauwerck vid Svenska barnboksinstitutet tar hjälp av stenografer och AI för att lära datorer att läsa Astrid Lindgrens manus.

MILJONER SIDOR

Riksarkivet förvarar 80 hyllmil handlingar och har över 200 miljoner inskannade sidor. Riksarkivet, ArkivDigital och Ancestry har över 200 miljoner skannade sidor från kyrkböcker. Kungliga biblioteket och universitetsbibliotek förvarar tusentals personarkiv med manus och brev.

Vid digital transkribering av handskrift måste datorn först identifiera vilka bildpixlar som sitter ihop i ord. Här är indata för utveckling av en sådan algoritm, där orden rutats in manuellt och betydelsen angetts. Utifrån många sidor med exempel lär sig datorn hitta ordbilden på andra ställen.

Sedan testar man om datorn känner igen samma ord på andra ställen. Blir det inte tillräckligt bra, eftersom handskrivna ord aldrig blir identiska, matar man in och tränar algoritmen med flera exempel. Exemplet ”finns” och ”Katla” på nästa uppslag är ett tydligt exempel. Att på så sätt lära datorer att läsa handskriven text kallas handskriven textigenkänning, eller HTR (Handwritten Text Recognition). Det är ett samlingsnamn på en rad algoritmer för avancerad bildbehandling, som löser delproblem inom området.

Forskningen kring transkribering av handskrifter till sökbar datakod har skjutit fart i takt med utvecklingen av artificiell intelligens, AI. Det innebär att datorer ”lär sig” lösa problem på sätt som liknar den mänskliga hjärnans, i så kallade neuronnätverk, genom parallella processer i kraftfulla datorer.

Projektet med Astrid Lindgrens ”krumelunsar”, som hon kallade sina stenograferade ord, går ut på att bygga algoritmer som kan läsa just hennes stil. Till ett system som klarar all stenografi är det däremot långt kvar.

Algoritmerna tas fram av forskare vid Uppsala universitet, som just nu fintrimmar den algoritm som känner igen olika ord.

– Stenografi är ljudhärmande och bygger på förenklingar. Alla sje-ljud skrivs till exempel likadant – stjärna och kärna blir samma ordbild. Det krävs mycket förfining om en dator ska göra rätt utifrån ordets sammanhang, säger Malin Nauwerck.

Nästa steg, som redan påbörjats, är att ta fram en algoritm som talar om var på sidan ordet finns, uttryckt i x- och y-koordinater – eftersom råmaterialet är en bild. När alla ordbilder tolkats till ord med kända positioner på sidan kan de sammanfogas till löpande text.

Projektet pågår i tre år till 2022.

Även Riksarkivet samarbetar med volontärer i ett pilotprojekt med polisrapporter i Göteborg från 1868 till 1902. Det är ett omfattande material med kanske 100 000 namn, platser och andra detaljer från utredningar.

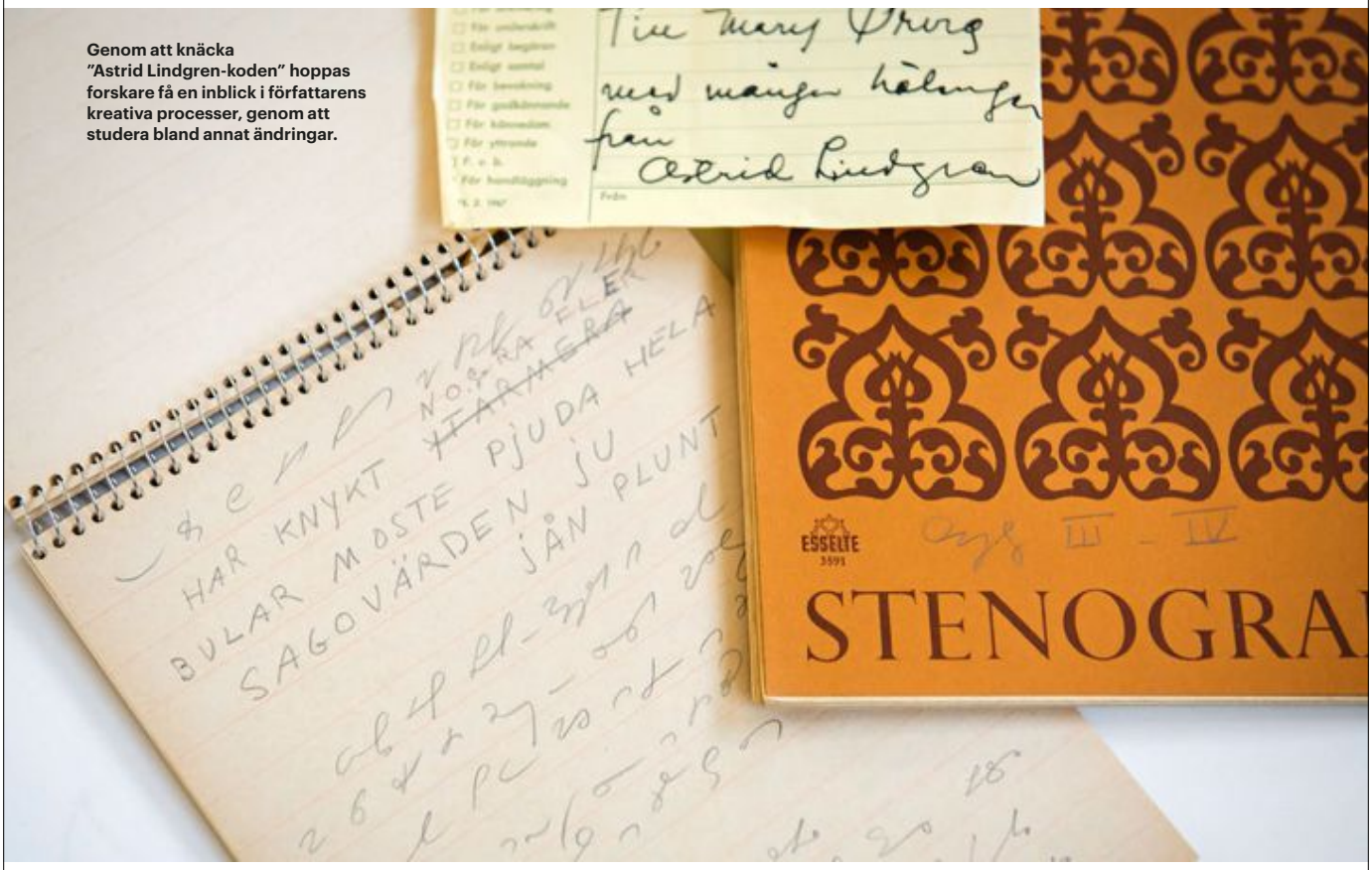
För att skapa underlag för transkribering tar arkivet hjälp av släktforskare och andra historiskt intresserade. De har

PROJEKTET GENUS OCH ARBETE/HISTORISKA INSTITUTIONEN/UPPSALA UNIVERSITET, EVA DALIN

KORT OM TEKNIKEN

Släktforskarnas och historikernas dröm om att kunna tolka gammal handskriven text digitalt kräver stor datorkraft. Handskrifterna läses in som bilder, vilka kopplas till stora mängder manuellt inmatad text som algoritmerna "tränar" med. Mycket forskning går ut på att minimera den mängden.

GETTY IMAGES



Genom att knäcka ”Astrid Lindgren-koden” hoppas forskare få en inblick i författarens kreativa processer, genom att studera bland annat ändringar.

läst utvalda sidor och skrivit ned dem i klartext. Båda parter ökar sina kunskaper och får ett slutresultat de har stor nytta av, framhåller projektledaren Karl-Magnus Johansson.

– Vi valde ett avgränsat material på 22 000 sidor. Rapporterna skrevs av ett tiotal tjänstemän, så vi har fått med olika handstilar. Sedan byggde vi en datormodell med 400 manuellt transkriberade uppslag som indata.

Sidorna skickades till Transkribus, en forskningsbaserad kommersiell aktör i Österrike som skapar algoritmer på beställning från forskare, arkiv och bibliotek världen över. Med hjälp av deras modell har Riksarkivet kunnat transkribera de 22 000 sidorna med 97 procents träffsäkerhet enligt företagets beräkning. En felprocent på under 10 procent brukar i dessa sammanhang betraktas som bra.

– Men när vi publicerar hela materialet vill vi att det ska vara så hundraprocentigt som möjligt. Så just nu läser våra externa deltagare igenom texterna och rättar fel.

Att hitta ord i en prydlig text är ganska enkelt. Men många äldre dokument är nötta, blekta eller vikta, text lyser igenom från baksidan, det finns bläckplumpar, andra fläckar och brännskador. Skribenten kan ha varit gammal och darrhänt. En text kan också innehålla förkortningar och den inskannade bildfilen kan ha för låg upplösning.

Det finns alltså många felkällor och kombinationerna av dem kan bli ändlösa. Många inskannade sidor måste därför ”tvättas”, vilket det också utvecklas algoritmer för inom HTR-forskningen.

Karl-Magnus Johansson leder Riksarkivets projekt med digital transkribering av handskrift.

Lasse Mårtensson, professor i nordiska språk vid Uppsala universitet, söker stildrag som skiljer medeltida skribenter åt.

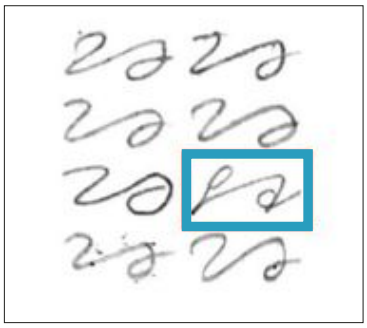
AI kan även hjälpa till med att lösa andra HTR-problem. En gigantisk utmaning är att identifiera vilka som skrev medeltida svenska dokument och när de skrevs, för att sätta in dem i en historisk kontext. Här är inte AI-systemets huvuduppgift att identifiera enskilda ord, utan stildrag.

– Det är svårare än vi hade trott. Skräp och skador på ett dokument kan få tolkningen att krascha helt, säger Lasse Mårtensson, professor i nordiska språk, som driver den språkliga delen av projektet vid Uppsala universitet.

– Tidiga diplom, från 1200-talet, är prydligt skrivna, medan de från senare medeltid kan vara rätt stökiga. De har ofta skrivits ganska snabbt och varje tecken kan se ut på flera olika sätt.

Därför studeras även andra egenskaper, som hur mycket texten lutar, hur långt vertikala staplar sticker upp eller

EVA DALIN, ASIMA CASIC, UPPSALA UNIVERSITET



Svårigheterna att lära en algoritm förstå stenografi illustreras av olika förekomster av ordet ”finns” i manuset till *Bröderna Lejonhjärta*. Det markerade ordet ser snarlikt ut, men är ”Katla”.

Handskrift i en polisrapport från Göteborg. Siffrorna anger radnummer och under bilden visas transkriberad text rad för rad. Siffran 4 anger sidnummer i dokumentet.

4–18 anledning tro, det Wallin hade
4–19 biträde i närheten, Nilsson för
4–20 att skydda sog från vidare miss-
4–21 handel, icke gjort något särdeles
4–22 motstånd; utan låtit fråntaga
4–23 sig penningpungen.
4–24 I anledning häraf hafva Polis-
4–25 konstaplarne N^{o} 78 Olsson och N^{o}
4–26 107 Hedlund den 29 sistl. November
4–27 på eft. m. anhållit ofvanbemälde
4–28 Wallin, hvilken då befanns innehafva

ner. Frågan är om det ens går att hitta ett enskilt mått som bergsäkert skiljer en skribent från en annan.

Drömmen för släktforskare, historiker, språkvetare, bibliotekarier och många andra är ett universalverktyg som kan ”översätta” vilken handskrift som helst – även gamla svårlästa, slarvigt skrivna på illa medfaret papper eller pergament – till digital klartext. Det är den heliga graalen, som dataforskare kallar det och försöker finna i sina algoritmer.

Det är en uppgift som kan tyckas omöjlig. Mängden handskrivet material i arkiv, museer och bibliotek är i det närmaste oändlig – Riksarkivet har 80 hyllmil tryckta, maskinskrivna och handskrivna dokument, som digitaliseras i hög takt. Till det kommer hundratals miljoner inskannade sidor på bland annat Kungliga biblioteket och släktforskarsajter.

– Det finns många som vill se den här typen av metoder utvecklas, där allt blir läsbart, till exempel släktforskare. Men för akademiska forskare som begränsar sig till en viss typ av material är det fantastiskt att kunna söka bland 5 000 sidor, det är de väldigt nöjda med, säger Anders Brun, pionjär inom HTR-forskningen i Sverige.

– Däremot är vi inte framme vid en sökmotor för all handskrift, men kanske halvvägs. Det finns tekniker för att söka i ganska stora material – man kan till exempel söka i en pdf numera. Men de metoder som vi har tagit fram fungerar bara på ett material i taget, en viss persons skrift eller en viss typ av dokument, säger Anders Brun.

Han och kollegan Anders Hast, professor på Institutionen för informationsteknologi vid Uppsala universitet, är inte alls säkra på att en algoritm som kan transkribera alla gamla handskrifter eller ett verktyg för sökning är möjliga att skapa.

– Jag försöker bygga algoritmer som kräver mindre datorkraft, är enkla att använda och som löser användarens specifika problem, till exempel att tolka en viss handstil snarare än alla, säger Anders Hast.

Anders Hast, professor på Institutionen för informationsteknologi vid Uppsala universitet, utvecklar algoritmer för digital läsning av bland andra Astrid Lindgrens stenografi.

Maria Ågren, professor i historia vid Uppsala universitet, forskar kring kvinnors förörjningsarbete, i bland annat digra domböcker.

Marcus Liwicki, professor i maskininlärning vid Luleå tekniska universitet, vill se mer samarbete mellan svenska forskare på området.

Anders Brun, forskare vid Institutionen för informationsteknologi vid Uppsala universitet, är en pionjär inom svensk forskning kring digital transkribering av handskrift.

Anders Brun driver ett projekt med nyckelordssökning, tillsammans med historieprofessorn Maria Ågren, också i Uppsala. Hennes forskarlag studerar kvinnors försörjningsarbete genom tiderna. Yrke anges sällan, men när kvinnor kallas till ting, som målsägare, åtalade eller vittnen, nämns ofta deras sysslor.

Därför vill de gärna kunna söka efter nyckelord, som olika sysselsättningar, i digra domböcker. Forskarna använder även suppliker, det vill säga besvärsskrifter från enskilda till överhetspersoner och myndigheter.

– Det är en allt hetare källa för historisk forskning internationellt. Om ett program kan identifiera var för oss intressant information finns på en sida, skulle vi kunna gå direkt till dessa avsnitt och till exempel titta på inledningarna, där de sökande presenterar sig, säger Maria Ågren.

Med tanke på de ofattbara mängderna handskrifter i världens arkiv är ett annat viktigt delområde att kunna trimma algoritmer så att de kräver mindre mängd manuellt transkriberad text. Med det jobbar både Anders Hast samt andra forskare vid Blekinge tekniska högskola och vid Luleå tekniska universitet.

Forskningen i Luleå leds av Marcus Liwicki, världskänd forskare inom maskininlärning, som leder eller deltar i forskningsprojekt världen runt. Han betonar vikten av att överbrygga klyftan mellan dataforskare och humanister, men även mellan olika lärosäten, för att få mer fart på utvecklingen.

Den centrala uppgiften är att lära datorerna vad de ska leta efter, utan mänsklig interaktion, anser han.

– Det är en av de riktigt stora utmaningarna. Men det kräver bra källmaterial, med tydlig text och högupplöst inskanning. ●

ANDERS HAST, RIKSARKIVET, UPPSALA UNIVERSITET, HARRY BRUN, LTU, RIKSARKIVET