FOLKRÖRELSEARKIVET
FÖR UPPSALA LÄN

Digitalisering med HTR: Från ”att läsa på skärmen” till datorer som kan läsa.

Kanske uppmärksammade en och annan att åklagaren Krister Peterssons vid Palmeutredningens avslutande presskonferens 10 juni 2020. framhöll de stora fördelarna med att utredningens omfattande arkiv hade digitaliserats och numera kunde ”läsas på skärmen”. Eller så måste man vara arkivarie för att se just den delen av presskonferensen som den viktiga nyheten. Den stora vinsten med det palmeutredningens digitala arkiv låg inte i att utredarna kan läsa på en skärm istället för på papper, utan att datorn också kan användas för att söka i texten. Digitaliseringen ger en maskinläsningsbar text. För utredarna måste detta ha underlättat återsökning med hjälp av olika sökord och därmed möjligheten att upptäcka bortglömda samband i den stora mängd information som samlats in under 34 års utredande.

För maskinskriven text, på skrivmaskin eller utskriven på en skrivare, är detta numera en enkel process. Med de enklaste skannrar kan vi skanna ett dokument på papper och vid inskanningen bestämma att vi vill ha en PDF-fil som både ger oss en bild av dokumentet för våra mänskliga ögonen och en binär textsträng, ”ettor och nollor”, inbäddad i PDF-dokumentet som datorn kan läsa. Textsträngen skapas vid inskanningen av en teknik som kallas OCR (Optical Character Recognition) och som vi är mest vana vid att möta som en streckkod på butiksvaror. En bild omtolkas av OCR till binär text och vidare till den uppsättning av siffror och bokstäver som datorn kan presentera för oss – på skärmen.

Datorn har betydligt svårare att förstå handskrifter (egentligen är det mest mjukvaran och algoritmerna som har det svårt, men för enkelhets skull så skriver jag ”datorn”). Vi kan nog tycka att olika personliga skrivstilar är svårlästa ibland, särskilt om de har skrivits för länge sedan, med en skrivstil som är annorlunda en den vi lärde oss i skolan. För en dator är detta ännu mer komplicerat och den måste få mycket hjälp från människor för att lära sig tolka rätt. Samtidigt utvecklas nu bildtolkning och artificiell intelligens för inlärning snabbt – inte minst i datorers förmåga till ansiktsigenkänning. OCR-tekniken för handskrifter har fått en särskild förkortning: HTR för Hand Written Text Recognition.

Kan vi få datorn att lära sig läsa handskrifter öppnas nya världar för oss. En digitaliserad handskrift idag är ofta verkligen ingenting mer än en bild som ”kan läsas på skärmen”. Datorn kan inte hitta någonting i bilden åt oss och vi får själva ägna möda åt att stava oss igenom svårlästa handstilar. Vi slipper kanske ta oss till ett avlägset arkiv och sitta i en läsesal, men det är hela vinsten. Men kan vi få datorn att söka i texten och att översätta svårtolkade handstilar till tydligare ”tryckbokstäver” går det att hantera också detta material på ett helt annat sätt. Vi kan hitta snabbare, läsa snabbare, skapa databaser och sammanställa uppgifter i texter med datorns hjälp. En 1500-talshandskrift blir inte svårtillgängligare än Aftonbladets nätupplaga. Om vi bara förstår orden den använder och ämnet som behandlas i texten. Den delen kan HTR inte hjälpa oss med.

Örjan