Nya ögon på Sveriges medeltida skrivare

Inledning

I projektet Nya ögon på Sveriges medeltida skrivare möts flera hundra år gamla handskrivna urkunder och modern datavetenskap i form av digital bildanalys, maskininlärning och datorlingvistik. Det övergripande syftet med projektet är att använda de nämnda metoderna för att utvinna information om de medeltida svenska skrivarnas skrift och språk. Bildanalys och maskininlärning används här för att analysera och kartlägga skrivarnas skrift, dvs. hur de specifikt utformade sina skrivtecken, medan metoder från datorlingvistiken används för att analysera den språkliga utformningen hos den text de producerade. Det material som analyseras är handskrivet material från det medeltida Sverige, dels s.k. diplom (medeltida brev), dels handskrifter (medeltida böcker). I de språkliga undersökningarna ligger fokus på svenskspråkigt material, men i de undersökningar som rör skriftens utformning används både svenskt och latinskt material.

Frågor som ställs är bland annat:

- Vilka egenskaper hos skrivtecknen är unika för en enskild individ jämfört med andra?
- Hur förändrar sig en enskild skrivares skrift över tid?
- Hur ser den språkliga variationen ut hos en enskild individ vid denna tid?

Tanken är att de metoder som utvecklas också ska kunna användas på annat handskrivet material, från andra tider och med annan proveniens.

Detta projekt är en del av ett större samarbete mellan Institutionen nordiska språk, Institutionen för lingvistik och filologi (datorlingvistik) och Institutionen för informationsteologi (bildanalys) vid Uppsala universitet som kallas From Quill to Bytes (Q2B). Information om detta samarbete finns här: https://www.it.uu.se/research/project/q2b

Materialet

Allt det material som analyseras i detta projekt är, som nämnts, handskrivet. I projektet analyserar vi de diplom som har bevarats till i dag, i den mån det finns tillgängligt i digital form. Dessa bevaras och är under utgivning vid Svenskt Diplomatarium (Riksarkivet). Diplom är medeltida brev, och de innehåller i första hand olika juridiska överenskommelser, t.ex. försäljning av mark eller lösöre, överlåtelser och liknande. Den stora fördelen med diplomen för detta forskningsprojekt är att de är daterade, och det betyder att de utgör fasta hållpunkter i undersökningar av skriftens utveckling över tid. Handskrifterna, de medeltida böckerna, är i regel odaterade.

En annan sak som präglar textframställningen och därmed också de handskrivna urkunderna från denna tid, och som är en följd av att allt producerades för hand, är att de bevarade handskrifterna sällan är originalversioner av texterna i fråga. Före tryckkonstens införande var det ända sättet att skapa en ny bok att skriva av en äldre förlaga för hand. De texter som ryms i de bevarade handskrifterna är därför i regel avskrifter, och ofta avskrifter i flera led. Vid avskriftsprocessen skapas också avvikelser i förhållande till den förlaga som skrivs av, delvis medvetna sådana, i form av olika större redaktionella ingrepp eller mindre stilistiska förändringar, dels omedvetna sådana, i form av misstag från skrivarens sida.

Till stor del faller det undersökta materialet inom den senare delen av medeltiden, dvs. slutet av 1300-talet och hela 1400-talet. Den skrift som förekommer då i Sverige, liksom i stort sett hela övriga Europa, tillhör den gotiska skrifttypen, och främst de kursiva varianterna (Cursiva Antiquor och Cursiva Recentior). Den senare delen av medeltiden var också en tid av mycket kraftig förändring av det svenska språkets struktur, på samtliga språkliga nivåer. Det svenska språket får sitt grundläggande uttal (förändring av stavelsestrukturen; vokaldansen), det gamla fyrkasussystemet börjar försvinna (en brytpunkt, åtminstone i Stockholmsområdet runt 1450) och språket tar in en mycket stor mängd lånord från lågtyskan.

Huvudområden

De frågor som ställs inom detta projekt, och den metodutveckling som drivs, syftar till stor del åt att analysera och kvantifiera de olika typer av variation som uppkommer som ett resultat av att allt skrivet material vid denna tid producerades för hand. Variationen som är i fokus faller inom tre olika kategorier:

1. Variation i skrivtecknens utformning.
2. Variation i den språkliga utformningen.
3. Variation i den textuella utformningen.

Inom samtliga dessa områden pågår arbete med olika digitala metoder, inom 1) bildanalys/maskininlärning, och inom 2) och 3) datorlingvistik. Målet är vidare att projektet ska generera ett större arbete, i monografiform, inom vart och ett av dessa områden ur ett filologiskt/språkvetenskapligt perspektiv. Dessa är under arbete. Nedan beskrivs huvudsakligen del 1, dvs. det som rör digital paleografi, eftersom det främst är vi har publicerat fram till nu. De övriga områdena kommer förhoppningsvis att generera publikationer under de kommande åren.

När det gäller skriftens utformning, kan man på ett övergripande plan närma sig ämnet från två skilda perspektiv:

1. Man kan söka efter och kvantifiera egenskaper knutna till det enskilda skrivtecknet. Till exempel, vilka egenskaper utmärker ’g’, ’h’, ’k’ osv. Det är till stor del här den traditionella paleografiska forskningen har befunnit sig, i det att man har undersökt hur de enskilda skrivtecknen har utvecklats över tid, och man har sökt formulera hur en viss skrivare har utformat vissa enskilda skrivtecken, för att särskilja denna från andra.
2. Man kan söka efter skriftegenskaper som återfinns i många olika skrivtecken och bildar ett genomgående mönster i skriften. Som exempel på sådana egenskaper kan nämnas skrivtecknens proportioner i förhållandet mellan höga skriftelement (staplar) och låga skriftelement (t.ex. minimer), och proportionerna mellan skrivtecknens höjd och bredd. Dessa egenskaper är svåra att mäta med ögonmått, åtminstone med någon grad av precision, medan de digitala metoderna erbjuder stora möjligheter.

Vi har gjort undersökningar inom båda dessa perspektiv, men med en övervikt för det senare. Nedan ges ett urval av de undersökningar vi har gjort. Denna översikt avslutas med en publikationsförteckning i urval.

Centroider

Ett exempel på perspektiv 1 ovan är en undersökning av handskriften Cod. Ups. C 64, innehållande Summula, av Laurentius av Vaksala. Ur denna handskrift samlade vi ihop exempel på ’a’:n, och sedan sammanställde datorn dessa till ett genomsnitts-’a’, en centroid. Bilden kan ses nedan (bild 1). Denna bild rymmer information om var samtliga ’a’:n sammanfaller (de mörkblåa fälten) samt var de varierar (röda och gula fält). Denna typ av undersökningar var vanliga i de tidiga försöken inom datorstödd paleografi. En förutsättning för att det ska gå att sammanfatta skrivtecken i genomsnittsformer är att de enskilda exemplen, graferna, är någorlunda lika varandra, dvs. skriften måste vara noggrant utförd. Så är fallet i C 64, och därför fungerade denna metod relativt bra här. Tyvärr är dock en stor del av det medeltida nordiska materialet inte av ett sådant utförande att denna metod är användbar i stor skala.

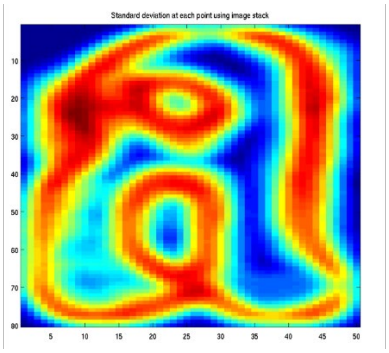

Bild 1. Centroid av ’a’.

Quill Features

Vi har gjort en undersökning som innebär att man följer hur skrivaren har fört pennan över skriftunderlaget, och hur han/hon har hållit i den. Denna undersökning faller inom perspektiv 2 ovan, dvs. den är inte knuten till enskilda skrivtecken, utan den riktar in sig på generella mönster. Den benämns ofta quill-metoden, på grund av den centrala roll som den medeltida pennan spelar i den, alltså fjäderpennan, *the quill*. På grund av fjäderpennans konstruktion, dess spetsning och dess snitt på toppen visar skriften utförd med den upp en variation mellan bredare och smalare streck, variation i bredd. Det är just den variationen som mäts med den här metoden: variationen mellan breda och smala streck, under beaktande av vinkeln som pennan har hållits i.

Det man således söker efter är den vinkel som en skrivare har hållit pennan i. Denna vinkel benämns *den regelbundna pennspetsvinkeln* (efter ”the habitual tip angle”), och den kan beräknas genom att man mäter vinkeln mellan de tunnaste strecken i skrivtecknen och baslinjen. De bredaste strecken uppkom när pennan drogs snett nedåt till höger och de tunnaste strecken uppkom när pennan rördes snett uppåt till höger. Och det är här pennspetsens vinkel mäts: vinkeln mellan det tunnaste strecket och baslinjen.

I bild 2 nedan är en av kvantifieringarna av quill-undersökningen, i form av histogram. Detta är histogram för två olika sidor i manuskriptet Cod. Ups. C 61, som vi undersökte med denna metod, Här är histogram för två sidor, producerade av olika skrivarhänder. Dimensionerna i histogrammen är pennvinkel (den lodräta dimensionen) respektive streckens bredd, den vågräta dimensionen.

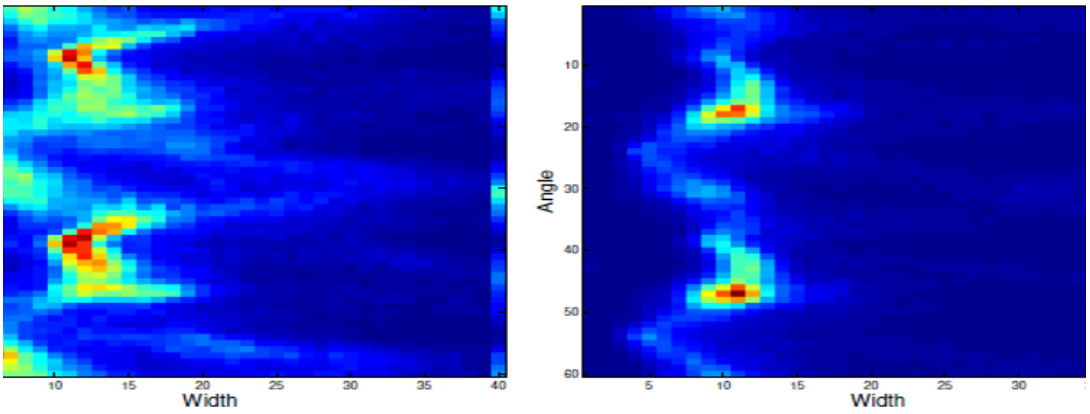

Bild 2. Histogram för sidor i C 61.

Framför allt är det den lodräta dimensionen som är av intresse, eftersom det är där pennvinkeln kan utläsas. Den vågräta dimensionen, variationen i streckens bredd, är främst en mätning av pennans bredd. Likväl måste denna variation mätas för att man ska kunna komma åt pennvinkeln. Man kan se att de färgade partierna, de i icke-mörkblått, sträcker sig över bredare partier i vågrät dimension i den vänstra bilden än i den högra, och det antyder att en penna har använts som kan skapa en stor skillnad mellan breda och smala streck. Sedan kan man också se att de rödfärgade partierna, dvs. där en viss vinkel har mött breda streck, är placerade vid olika höjd i den lodräta dimensionen. Det betyder att dessa två skrivare har hållit pennan i olika vinkel. När man sedan tittar på sidorna bakom histogrammen så ser skrivarhänderna verkligen olika ut. Detta är handen bakom det vänstra histogrammet:

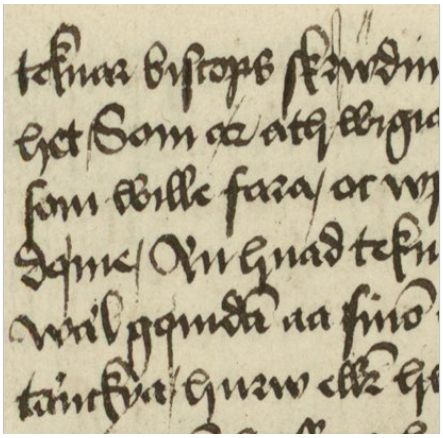

Bild 3. Skrivarhand, vänstra histogrammet.

Detta är handen bakom det högra histogrammet:

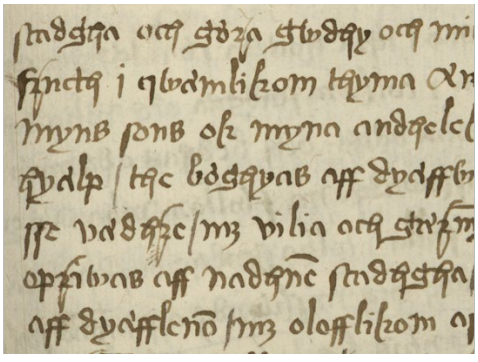

Bild 4. Skrivarhand, högra histogrammet.

En annan visualisering av resultaten från denna undersökning visas nedan (bild 5), en s.k. scatter plot. Var och en av prickarna i denna plot är en sida från manuskriptet C 61, och prickens placering i bilden baserar sig på mätningen av quill-egenskaperna. Här kan vi därmed se hur de olika sidorna och skrivarhänderna i handskriften är relaterade till varandra avseende quill-egenskaperna. Färgerna är de olika skrivarhänderna, färgade av oss i förväg, baserat på tidigare forskning. De gula sidorna har producerats av handen på bild 4 ovan. De ligger tydligt för sig själva, medan de övriga ligger i ett kluster, dock regelbundet inom detta.

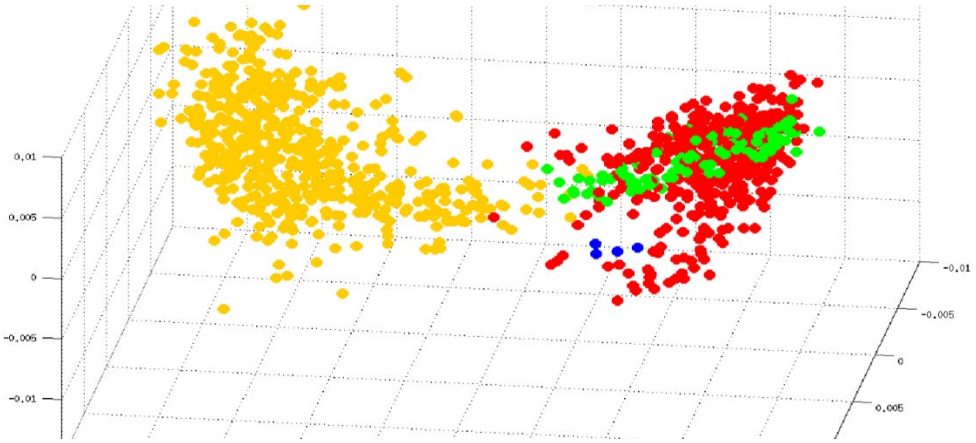

Bild 5. Scatter plot, C 61.

Vi har sedan utvidgat denna undersökning till ett nytt material, nämligen 12 000 medeltida svenska diplom (det antal vi har tillgängligt i digital form). Svårigheterna här i förhållande till C 61 är flera. Quill-undersökningen på C 61 var för det första mindre i omfång (ca 1100 bilder gentemot ca 12 000), och för det andra var det fråga om en enskild handskrift. C 61 består visserligen av flera från början självständiga delar, men man kan ändå anta att de har tillkommit nära varandra i tid. Vidare är det fråga om ett begränsat antal skrivarhänder (fyra alternativt fem), och de flesta av dem har skrivit många sidor. Slutligen har handskriftssidorna samma format och ungefär samma omfång. Bilderna har därmed också samma format, och de har tagits vid samma tillfälle. Bilderna av handskriftssidorna är alltså helt jämförbara bildtekniskt, avseende de data som kan utvinnas.

Diplomen i SDHK sträcker sig mellan 1100-talet till ca 1500, vilket innebär att det är ett långt tidsspann mellan själva urkunderna. Sedan är det fråga om väldigt många skrivare. Och tittar man på bild 5 ovan, så ser man att flera enskilda handskriftssidor, skrivna av olika skrivare, har hamnat nära varandra. Det som gör att man kan se att metoden ändå fångar relevanta egenskaper är att mängden exempel fördelar sig regelbundet. Undersökningen av diplomen är under arbete.

Dessa undersökningar har utförts av mig själv, Fredrik Wahlberg och Anders Brun. En artikel som den har genererat finns här: http://www.sol.lu.se/uploads/media/ANF_130__2015_.pdf

Mätning av skriftproportioner

I en annan undersökning ligger fokus på proportionerna mellan höga skriftelement, till exempel staplar (höga komponenterna i ’b’, ’h’, ’d’ etc.), element i medelhöjd, t.ex. minimer (det vill säga korta vertikala streck, t.ex. den låga komponenten i ’h’, vidare ’i’, ’m’, ’n’ etc.)

och öglor i medelhöjd (t.ex. den låga komponenterna i ’b’ och ’d’, vidare enrummigt ’a’, ’o’ etc.), och låga element, ibland kallat underhäng, som sträcker sig under raden (t.ex. i ’p’, ’q’ etc.). På bild 6 ser vi en illustration av detta.

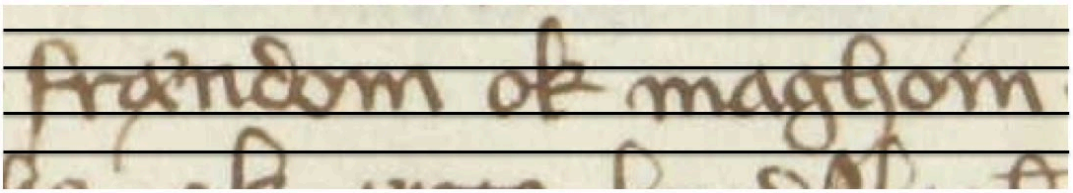

Bild 6. Den tredelade indelningen av skriftytan.

Vi har här ett mittfält som innehåller elementen i mittenhöjd och sedan element som sträcker sig ovanför och under detta. Observera här att det här är en schematisk bild, som bara visar principen, och linjerna är helt raka och följer inte skriften exakt. Denna specifika aspekt av skriften har länge varit i fokus för intresse för paleografin, men eftersom dessa egenskaper är mycket svåra att mäta utan datorer, har det varit svårt att avgöra vilken betydelse den har för skriftens kronologi och individualitet.

Den tekniska huvudfrågan här är att datorn ska identifiera mittfältet, som omsluter elementen i medelhöjd. Därefter identifierar och mäter datorn de komponenter som sträcker sig från mittfältet. I bild 7 kan vi se hur denna mätprocess ser ut:

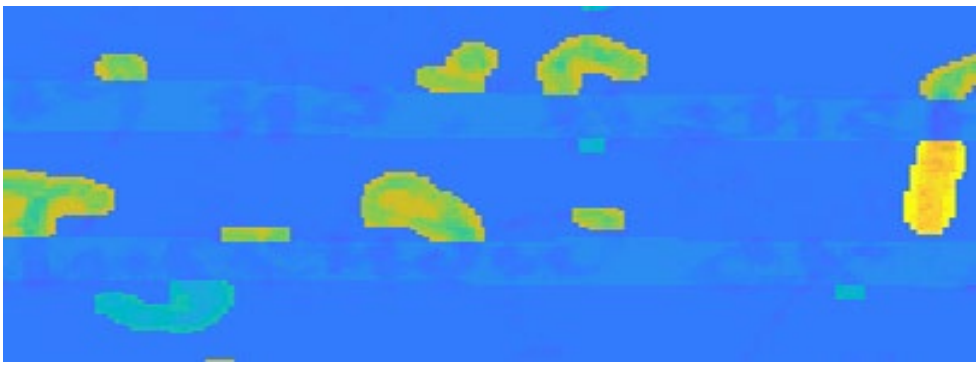

Bild 7. Mätning av de tre skriftytorna.

Det finns ett ljusblått område i mitten, och man kan urskilja svaga bokstäver inom detta. Det är mittfältet, såsom det har identifierats av datorn. Som ni kan se är mittfältet inte rakt, som i föregående schematiska bild, utan den går något upp och ner, som fallet är i verklig, handskriven skrift. Sedan stiger de höga komponenterna över mittfältet, markerade här i gult, sålunda registrerade och uppmätta. De låga elementen, som sträcker sig under mittfältet, är markerade i ljusare blått i bilden.

I bild 8 visas resultatet av mätningen. I bilden är varje prick ett diplom, och positionen i bilden är baserad på proportionerna mellan de höga elementen, mittfältet och de låga elementen. En skrivare vid namn Johan Sigmundsson har skrivit tre självattribuerade diplom, och de har blivit färgade i rött i bilden. Som framgår av bilden ser man bara en enda stor röd punkt, och det beror på att de ligger nästan ovanpå varandra. Dessa diplom kommer därmed mycket nära varandra när det gäller de aktuella skriftproportionerna. Dessa egenskaper förefaller sålunda åtminstone delvis vara betingade av individualitet. Alla självattribuerade diplom av

samma skrivare, vilka vi har använt i utvärderingssyfte, visar dock inte så här god överensstämmelse, men de ligger i regel samlat.

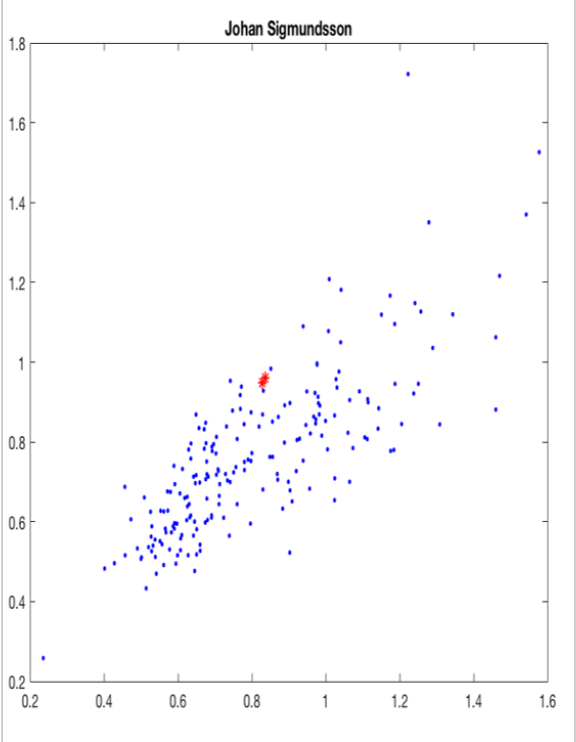

Bild 8. Kvantifiering av undersökningen av skriftproportionerna.

Denna undersökning har utförts av mig själv, Anders Hast och Ekta Vats. Vi har hållit en presentation om en pilotundersökning, och den fullskaliga undersökningen är under arbete.

Datering

Vi har också gjort en undersökning där vi har använt skrivtecknens form som utgångspunkt för datering, dvs. bestämning av dokumentens tillkomsttid. Fokus ligger här inte på enskilda skrivtecken, utan generella mönster (dvs. undersökningen faller inom perspektiv 2 ovan). Försöket utgör början på en ambition att hitta kronologiska hållpunkter i skrivtecknens utformning. Det finns naturligtvis en omfattande tidigare forskning om paleografi i dateringssyfte, och det är viktigt att följa upp det som man har kommit fram till där. Men detta är ett försök att finna nya kronologiska företeelser, sådana som vi ännu inte har uppfattat, eller åtminstone inte formulerat tydligt ännu. Det har vi gjort genom att låta datorn själv, genom maskininlärning, söka efter mönster i skriften och sedan relatera dessa data till produktionsår. Men det svåra sedan är att komma åt vad datorn har mätt. Häri ligger vår stora utmaning, och det är något som vi håller på att arbeta med nu.

Processen går till i följande steg:

1) Skrivtecknens konturer mäts (på samtliga 12 000 diplom). Konturerna är därmed den enda faktorn som egenskaperna extraheras från. Det är utifrån pixlarna i bilden av konturerna som skrivtecknens egenskaper utvinns. Mätningen av konturerna illustreras i bild 9.

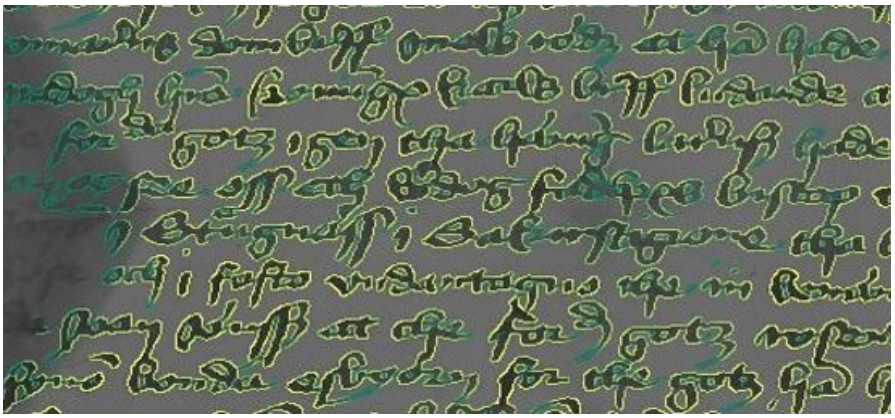

Bild 9. Mätning av skriftens konturer.

2) Egenskaper extraheras från konturerna av skrivtecknen. Extraheringen av egenskaperna görs genom att en mall placeras över varje pixel i bilden av skriftens kontur. Datorn läser av och kartlägger omgivningen kring denna pixel, och utvinner mönster ur detta. Bild 10 illustrerar detta.

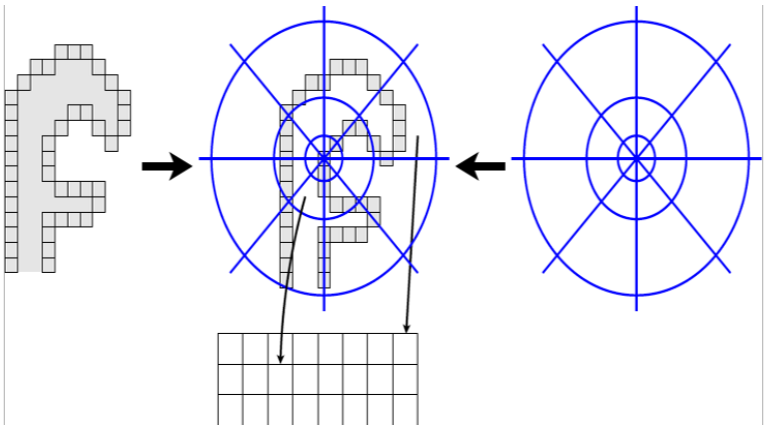

Bild 10. Egenskaper extraheras.

3) De utvunna egenskaperna relateras till diplomens produktionsår. Det är gjort på ett urval (5% av korpusen). Datorn sätter således in de utvunna egenskaperna i ett kronologiskt perspektiv. Det ska observeras att urvalet diplom som datorn har tränat på är jämt fördelat kronologiskt, dvs. lika många diplom har valts från 1400-talet som från 1300-talet (se bild 11 nedan).

4) Datorn använder den extraherade informationen på diplom där produktionsåret inte ges till datorn, och en uppskattning av produktionsår görs. I bild 11 relateras de diplom som användes till träningsprocessen i steg 3 ovan till de självständiga dateringar som datorn gör i detta steg:

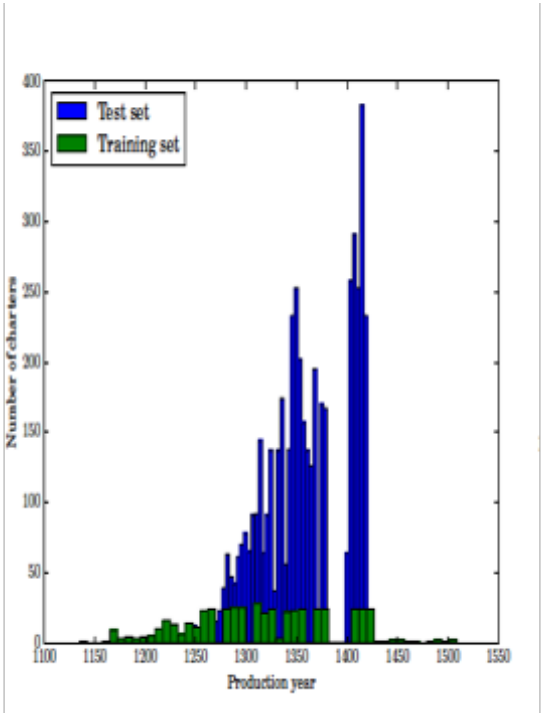

Bild 11. Dateringarna jämte träningsmaterialet.

Eftersom diplomen är daterade kan den datering som datorn gör kontrolleras. Dateringarna blev relativt bra, och de ligger vanligtvis inom intervallet plus/minus ca 20 år från det korrekta årtalet. Slutsatsen här måste vara att datorn har upptäckt några egenskaper som verkar vara relevanta ur ett kronologiskt perspektiv, och som måste ha att göra med skrivtecknens form, eftersom det var utifrån den som mätningen utgick. Det som kvarstår är en utvärdering av de egenskaper som har varit utslagsgivande, och det är ett arbete som pågår nu.

Detta arbete har utförts av Fredrik Wahlberg, mig själv och Anders Brun. En publikation från denna undersökning finns här:
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7490092

Avslutning

Detta är ett litet urval av de undersökningar som har gjorts inom ramen för detta projekt. Som nämnts ovan ligger fokus i beskrivningen på de undersökningar inom digital paleografi som har utförts inom ramen för detta projekt. Det ska betonas att vi har arbetat betydligt bredare än vad ovanstående beskrivning ger vid handen. Ett område som är mycket centralt, och som har varit i fokus i flera publikationer, är att reducera bort smuts, märken etc. från bilderna, för att datorn i mätningarna enbart ska processa skriften och inget annat.

Ett centralt område som tillkommer, och som har utvecklats i hög utsträckning under projektets senare tid, är de digitala applikationerna på medeltida text. Målet är att dessa undersökningar ska generera publikationer under kommande år. Materialet är också här medeltida svenskt material, och i fokus här är, liksom i paleografiundersökningarna, de medeltida svenska skrivarna, undersökta med datorstödda metoder som tillåter behandling av ett stort material.

Publikationer

Ett urval publikationer från projektet ges nedan (i omvänd kronologisk ordning):

In Search of the Scribe: Letter Spotting as a Tool for Identifying Scribes in Large Handwritten Text Corpora. Mårtensson, L. Vats, A. Hast, A. Fornés. Human IT, Vol. 14, No. 2. 2019.

Radial Line Fourier descriptor for historical handwritten text representation.
A. Hast, E. Vats. Journal of WSCG, ISSN 1213-6972, Vol. 26, No. 1, Journal. pp. 31-40. 2018.

Making Large Collections of Handwritten Material Easily Accessible and Searchable
A. Hast, P. Cullhed, E. Vats, M. Abrate. Proceedings of the Italian Research Conference on Digital Libraries (IRCDL), Full Paper. 2019.

Finding Logo and Seal in Historical Document Images - An Object Detection based Approach Sukalpa Chanda, Prashant Kumar Prasad, Anders Hast, Anders Brun, Lasse Martensson and Umapada Pal. The 5th Asian Conference on Pattern Recognition (ACPR 2019). 26-29 November 2019, Auckland, New Zealand.

Learning Surrogate Models of Document Image Quality Metrics for Automated Document Image Processing. P. Singh, E. Vats, A. Hast. 13th IAPR International Workshop on Document Analysis Systems (DAS), Full Paper. pp. 1-6. 2018.

TexT - Text Extractor Tool for Handwritten Document Transcription and Annotation.
A. Hast, P. Cullhed, E. Vats. 14:th Italian Research Conference on Digital Libraries, IRCDL, Full Paper.. 2018.

Automatic Document Image Binarization using Bayesian Optimization. E. Vats, A. Hast, P. Singh. The 4th International Workshop on Historical Document Imaging and Processing (HIP 2017), in conjunction with ICDAR 2017, Full Paper. 2017.

On-the-fly Historical Handwritten Text Annotation. E. Vats, A. Hast. 1st International Workshop on Human-Document Interaction (HDI 2017), in conjunction with ICDAR 2017, Full Paper. 2017.

Automatic Scribe Attribution for Medieval Manuscripts. M. Dahllöf. Digital Medievalist 11(1): 6. 2018.

Clustering Writing Components from Medieval Manuscripts. M. Dahllöf. I: Michael Piotrowski (red.) COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018, Lausanne. 11-13. 2018.

Dahllöf, Mats. 2019. “Clustering writing components from medieval manuscripts.” I Michael Piotrowski (red.) Proceedings of the Workshop on Computational Methods in the Humanities 2018, 23-32.

Paleografiska egenskaper ur ett digitalt perspektiv. L. Mårtensson, F. Wahlberg, A. Brun. Selskab for østnordisk filologi 2. Universitets-Jubilæets danske Samfund, University Press of Southern Denmark. 2017.

Large scale continuous dating of medieval scribes using a combined image and language model. Med Fredrik Wahlberg och Anders Brun. I: Proceedings of the 12th IAPR Workshop on Document Analysis Systems (DAS 2016).

Digital Palaeography and the Old Swedish Script. The Quill Feature Method as a Tool for Scribal Attribution. Med Fredrik Wahlberg and Anders Brun. I: Arkiv för nordisk filologi 130/2015.

Large Scale Style Based Dating of Medieval Manuscripts. F. Wahlberg, L. Mårtensson, A. Brun. I: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing (HIP'15).

Scribal Attribution using a Novel 3-D Quill-Curvature Feature Histogram. F. Wahlberg, L. Mårtensson, A. Brun. I: Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR-2014).

Data Mining Medieval Documents by Word Spotting. F. Wahlberg, M. Dahllöf, L. Mårtensson, A. Brun. I: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing.