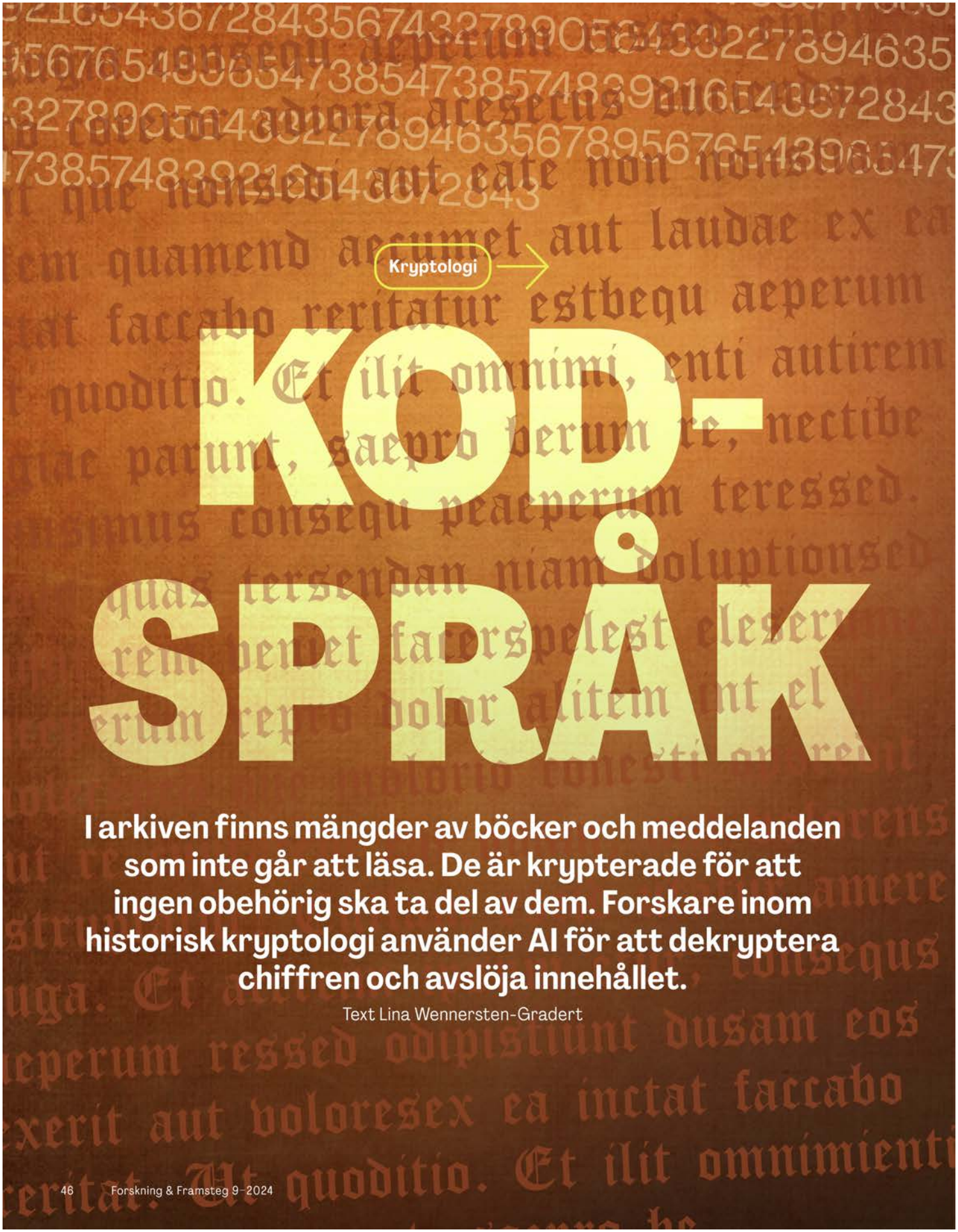


Kryptologi

KOD-SPRÅK

I arkiven finns mängder av böcker och meddelanden som inte går att läsa. De är krypterade för att ingen obehörig ska ta del av dem. Forskare inom historisk kryptologi använder AI för att dekryptera chiffren och avslöja innehållet.

Text Lina Wennersten-Gradert

En rörig och omisskännligt brittisk blandning av arkitekturelement har sammanfogats i grönskan vid en sjö. Det skulle kunna vara en scen för ett kostymdrama eller en Agatha Christie-deckare, med inslag av rävjakt och överdådiga tebjudningar. Men godset sju mil nordväst om London är ikoniskt av en helt annan anledning. Det var här som Alan Turing och hans kodknäckargrupp överlistade Enigma, tyskarnas berömda krypteringsmaskin. Kanske inte precis inne i själva herrgården, men hela Bletchley Park, som efter hand fylldes på med byggnader, är genom otaliga böcker, filmer och tv-serier förknippat med den brittiska underrättelsetjänsten i allmänhet och med kodknäckarna i synnerhet. MI6 köpte lantegendomen 1938 och där inhystes GHCQ, Government code and cypher school. Året därpå fick en växande skara matematiker, schackspelare och korsordsspecialister ärva Polens dechiffreringsarbete, vilket gav dem en bra start, men att avkoda tyskarnas meddelanden tog tid.

Enigma kombinerade mekanik och elektronik för att kryptera information på ett ständigt förändrat sätt. Varje bokstav kunde kodas annorlunda för varje gång den användes. Ett A kunde alltså först betyda H, sedan R och därefter O. Enigma kunde ställas in på ett nästan ofattbart antal olika sätt. För den som inte hade kodnyckeln och en maskin för att omvandla den krypterade texten till ett läsligt meddelande var det som att hitta en nål i universum. Om den som sände informationen gjorde någon miss, så att krypteringen inte ändrades varje gång en bokstav förekom, blev det något lättare. Men även om kodknäckarna hittade nålen så kastades den ut i universum igen nästa dag när kodnyckeln byttes. Som avgörande stöd i detta tröstlösa sisyfosarbete konstruerade Alan Turing den första prototypen till The bombe, en sorts elektromekanisk maskin för att leta kodnycklar. Med hjälp av bombe-maskiner och en beslagtagen Enigma, och tack vare småslarv på sändarsidan, kunde de allierade tolka motståndarens meddelanden. På Bletchley Park konstruerades även en av världens första datorer, Colossus, som användes för att avkoda meddelanden skapade av en annan av tyskarnas krypteringsmaskiner: Lorenz SZ40/42.

– Utan dessa framgångar kanske tyskarna hade vunnit, säger Beáta Megyesi, professor i datorlingvistik vid Stockholms universitet och en spindel i nätet inom forskningsfältet historisk kryptologi.

MÅNGA BETYDLIGT ENKLARE krypteringar är däremot ännu olösta. Runt om i världens arkiv finns tusentals texter skrivna med chiffer. Militär-

INGMARIE ANDERSSON/STOCKHOLMS UNIVERSITET

Beáta Megyesi är professor i datorlingvistik och arbetar med historisk kryptologi.

strategisk korrespondens, statshemligheter och underrättelserapporter, men också skrifter från hemliga sällskap, vetenskapliga redogörelser och privata kärleksbrev. De är skrivna i olika tider och vitt skilda sammanhang – det gemensamma är att avsändaren har lagt sig vinn om att ingen utomstående ska kunna förstå innehållet.

– Så länge som vi har haft skriftspråk har det funnits behov av att hemlighålla information. Och ju fler som lärde sig läsa, desto större blev behovet av att kryptera meddelanden, säger Beáta Megyesi.

Hon leder det tvärvetenskapliga forskningsprojektet Decrypt och var initiativtagare till den årliga internationella forskningskonferensen Histocrypt, som tar upp allt som rör kryptering, från begynnelsen, fram till och med kalla kriget. Vilket som är det äldsta bevarade chiffret är svårt att fastställa, berättar hon.

– Det finns många inskriptioner skrivna på sten, ben eller papper som vi ännu inte kan tolka. Vi vet inte om de är krypterade, om de är skrivna på ett okänt språk, eller om de är artificiella och påhittade, vilket innebär att vi kanske aldrig kommer att kunna förstå dem. När det gäller tidiga chiffer är Julius Caesar-chiffret ett av de mest kända exemplen.

DEN ROMERSKE KEJSAREN Julius Ceasar (100–44 f.Kr.) skriver i sina kommentarer om det galliska kriget (*Commentarii De Bello Gallico)* om hur krypterade meddelanden används för att fienden inte ska kunna läsa dem. Men om någon i fiendestyrkan hade en stund över skulle det gå ganska

fort att avkoda budskapet i en Ceasar-chiffrerad text. Det är nämligen den enklaste formen av kryptering – man byter ut bokstäverna genom att flytta ett visst antal steg framåt i alfabetet. Den som vill skicka ett kodat meddelande ser då till att mottagaren får veta kodnyckeln, antingen genom att man har kommit överens om den innan, eller genom att den skickas separat från meddelandet. Om Cicero får en lapp från Ceasar med bokstäverna HFYVK IOKR, och vet att kodnyckeln är 5 steg, så kommer han utan problem räkna sig fram i det latinska alfabetet och förstå att han ska fånga dagen (*carpe diem*). Detta är ett exempel på ett så kallat substitutionschiffer, att ett tecken ersätts med ett annat. Den andra varianten är transpositionschiffer där ordningen på bokstäverna kastas om. Ett enkelt sätt att kryptera *Carpe diem* genom transposition skulle då kunna vara att först ställa upp bokstäverna så här:

CAR
PED
IEM

Och sedan skriva dem nerifrån och upp i en sammanhängande följd: IPCEEAMDR.

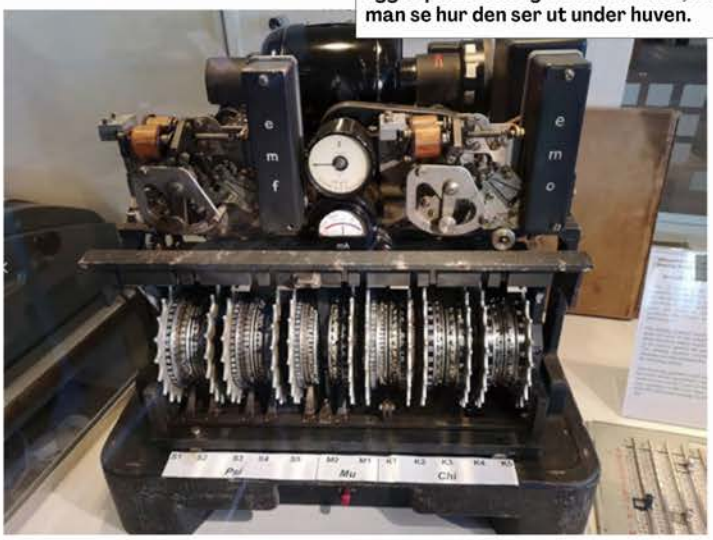
UNDER DEN MUSLIMSKA guldåldern, som inleddes omkring år 750, gjordes stora framsteg på chifferområdet, både i kodandet och avkodandet. Den arabiske matematikern Al-Kindi skapade redan på 800-talet en metod för att knäcka substitutionschiffer.

– Araberna utvecklade ganska avancerade tekniker och det var till exempel vanligt att använda chiffer för att dölja information om hur stor egendom folk hade, säger Beáta Megyesi.

De insåg till exempel att det gick att knäcka en kod genom att koncentrera sig på de vanligaste bokstäverna. Räknar du hur många gånger varje tecken förekommer i ett chiffer så kan du räkna ut tecknets andel av totalen. Det kan sedan jämföras med hur vanlig en viss bokstav är i det aktuella språket. Om man ska avkoda ett chiffer som utgår från svenska och upptäcker att X är det vanligaste tecknet, kan man misstänka att X ska läsas som E, eftersom det är den vanligaste bokstaven i svenskan. Den här säkerhetsluckan kan täppas till genom att ge de vanligaste bokstäverna flera olika tecken. Och som ytterligare förstärkning kan vissa ord, namn och platser få egna tecken.

I skriftspråkets tidiga historia användes i allmänhet inte mellanrum mellan orden – det kom först under 700–900-talet, även om romarna ibland satte punkter mellan orden. Att återgå till att skriva utan skiljetecken är ett vanligt sätt att göra chiffret säkrare, och det är mer regel än undantag att de historiska chiffer som hittas består av en enda lång teckensekvens. Och med tiden uppfinns nya sätt att göra krypteringen svårare att knäcka. Under 1300-talet börjar alfabetets tecken ersättas med symboler hämtade från till exempel astrologi och alkemi, och från 1400-talet och framåt tar siffror över mer och mer. Eftersom det behövs en siffra för varje alfabetiskt tecken, och sedan eventuellt siffror för några ord och namn så räcker det inte med talen 0–9.

Kodmaskinen Lorenz SZ42 användes av tyskarna på högsta ledningsnivå. På National Museum of Computing, som ligger på Bletchley Park-området, kan man se hur den ser ut under huven.



TED COLES

– Du måste ha mycket fler! Så man har tillämpat minst tvåsiffriga koder, och för att göra det svårare kunde man ha en tvåsiffrig kod för alfabetiska tecken, men en tresiffrig kod för ord. När ordgränserna inte markeras blir det bara en enda lång siffersekvens. Ett sådant chiffer blir ganska säkert, säger Beáta Megyesi.

Men krypteringstekniken fortsätter ändå att utvecklas under 1500–1700-talen och kodnycklarna blir till tjocka kodböcker.

– Du kan föreställa dig det som ett gigantiskt lexikon där varje ord i språket har fått sin egen kod. Även böjda former av orden, så ”gå”, ”går” och ”gick” skulle få tre olika koder.

Bara den som har koden ska kunna läsa texten, och den finns förhoppningsvis i tryggt förvar hos mottagaren. När det gäller historiska chiffer är mottagaren död sedan länge, men i arkiven återfinns en del kodnycklar.

– Sällan hittar vi brev och nycklar på samma ställe, eftersom de förvarades åtskilda på olika platser, säger Beáta Megyesi.

DEN GRANNLAGA UPPGIFTEN är att para ihop kodnycklarna med rätt chiffer, och framför allt att avkoda innehållet när nyckeln saknas, vilket den oftast gör. Medan kryptologerna på Bletchley Park var tvungna att uppfinna maskiner och datorer, kan forskarna inom historisk kryptologi utgå från dagens betydligt mer avancerade datorer för att knäcka chiffren.

– De är extremt snabba på att beräkna saker och ting. Ett av de mest utmanande problemen, både för datorerna och för oss, är att identifiera enskilda koder i en siffersekvens, alltså att avgöra var gränserna mellan koder går. Beáta Megyesi och hennes team har skapat en databas som heter det den ska göra: Decode. Där matas tusentals historiska chiffer och nycklar in.

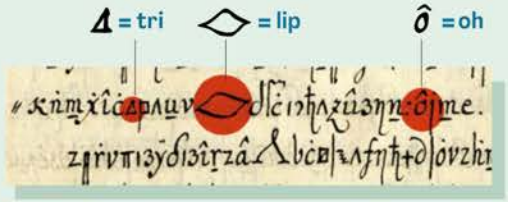
Det första steget är att transkribera texten, alltså att omvandla digitaliserade bilder av manuskript till något som kan läsas i och av en dator. Till det används AI-verktyg som måste kunna känna igen många olika typer av symboler och tecken.

Så blir historiska chiffer läsbara

AV JOHAN JARNESTAD OCH LINA WENNERSTEN-GRADERT

1 DIGITALISERING OCH TRANSKRIBERING

När ett chiffer har hittats är det första steget att digitalisera det. Bilder av sidorna läggs in i databasen Decode tillsammans med metadata, alltså information om chiffret. Sedan behöver bilden eller bilderna av manuskriptet omsättas till datorläsbar text. Det kan göras helt manuellt eller halvautomatiskt med hjälp av ett AI-verktyg för handskriftsläsning och symbolöversättning: Transcript tool. Här är tre exempel på transkribering av symboler i det så kallade Copiale-chiffret:

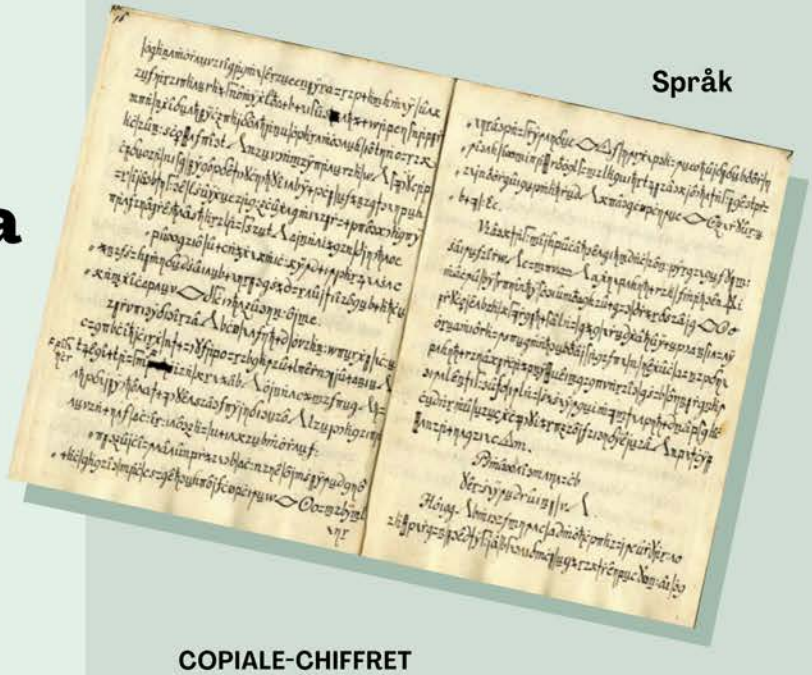


2 ANALYS

För att identifiera vilket språk som chiffret utgår från används språkmodeller baserade på historiska språkdata. Uppbyggnaden av den krypterade texten kan då jämföras med olika språks uppbyggnad. En svårighet är att stavning och vokabulär varierar mycket över tid, vilket AI-verktygens algoritmer måste ta hänsyn till. Dessutom kan chiffrens ursprungstexter växla mellan flera språk – några ord kan till exempel vara skrivna på latin och andra på franska, medan huvudspråket kan vara ett tredje. Forskarna använder verktyg för att göra frekvensanalys, alltså att räkna hur ofta olika tecken förekommer, för att kunna jämföra med hur ofta en viss bokstav förekommer i olika språk. I Copiale-chiffret var exempelvis ∧ vanligast.

3 DEKRYPTERING

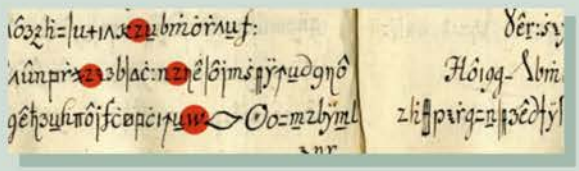
När forskarna ska försöka dekryptera ett historiskt chiffer är utgångspunkten att det är ett substitutionschiffer, eftersom det är vanligast, alltså att en bokstav eller ett ord har ersatts med vissa tecken, siffror eller symboler. Men substitutionschiffren kan vara uppbyggda på sinsemellan olika sätt. Forskarna har skapat verktyg för att gruppera chiffer som finns i databasen och som liknar varandra. Med hjälp av verktyget CrypTool 2 kan till exempel de flesta krypterade texter från Vatikanens arkiv avkodas.



COPIALE-CHIFFRET

Bakgrunden till att de tvärvetenskapliga internationella forskningsprojekten Decode och Decrypt föddes var avkodningen av Copiale-chiffret. På första uppslaget i manuskriptet stod det: ”Philipp 1866”, vilket antogs vara en ägarmärkning, och på sista sidan stod det ”Copiales 3”.

Allt annat i dokumentet var krypterat och ursprungsspråket i den kodade texten förmodades vara tyska. Chiffret visade sig vara ett så kallat homofoniskt substitutionschiffer, alltså att vissa bokstäver kodades med flera olika symboler. Latinska bokstäver användes som skiljetecken.

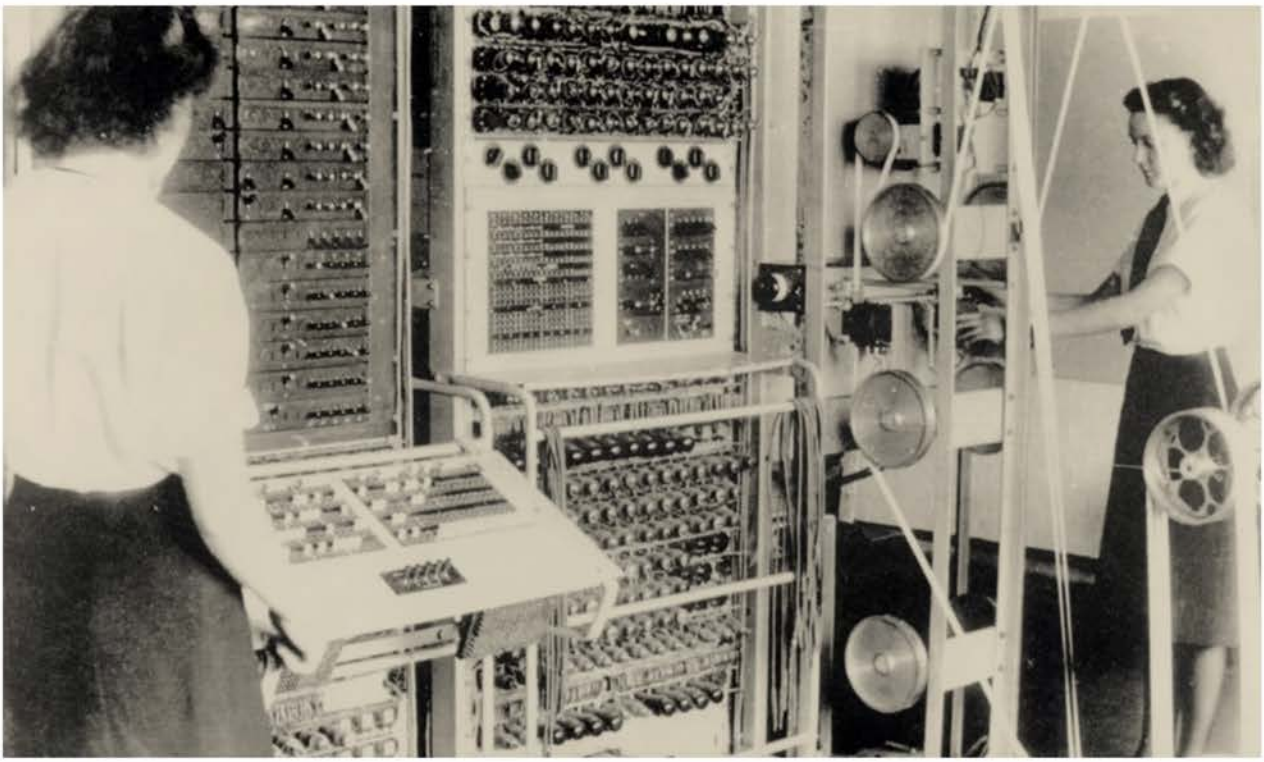


De tre forskarna Kevin Knight, Beáta Megyesi och Christiane Schaefer lyckades knäcka koden 2011. Manuskriptet daterades till omkring 1730–1760 och härrör från en hemlig orden: okulisterna.

CHIFFERNYCKEL

A	I	R	Y
B	J	S	Z
C	K	T	SCH
D	L	U	SS
E	M	Ü	ST
F	N	V	CH
G	O	W	Ä
H	P	X	

Språk



– AI i dag baseras i vanliga fall på extremt stora datamängder – om man dammsuger hela internets data får vi bra modeller som kallas LLM:s, *large language models*. GPT är ett exempel på en sådan stor språkmodell. För mig är det en utmaning att bygga AI-modeller som kan lära sig av små mängder data och anpassa sig till forskarens specifika behov, säger Beáta Megyesi.

AI:n kan till exempel nyttjas för att strukturera symboler och föreslå vilka symboler som liknar varandra. Sedan kan forskarna ta ställning till förslagen, ge återkoppling till AI:n och be om andra exempel. Forskarna kan också bygga språkresurser utifrån historiska texter så att de kan testa chifferlösningar mot flera olika språk. Ett steg är också att avgöra om en text verkligen är skriven på chiffer eller bara på ett okänt språk. För att avgöra det görs lingvistiska analyser med AI:s hjälp. Då går det att se strukturella likheter mellan en funnen text och ett visst språk. Ytterligare en uppgift för AI är att bistå med själva dechiffreringen.

– För kryptologer är det en utmaning att dechiffrera ett dolt budskap i ett historiskt chiffer – de ser det som ett slags sudoku. Men chiffer är också ett diversifierat sätt att kommunicera. För mig handlar det därför om metodutveckling i såväl humaniora som AI: Hur kan vi bygga system och AI-modeller som kan hjälpa oss att tolka historisk text?

Beáta Megyesi har vigt sin forskarkarriär åt den frågan och har under resans gång fått tillgång till Vatikanens annars så hemliga arkiv. Där har hon sökt efter chiffrerad korrespondens från 1300-talet och framåt och det var också i det påvliga arkivet som hon hittade det så kallade Borg-chiffret, daterat till 1630–1650-talet. Manuskriptet är på cirka 400 sidor och visade sig redogöra för dåtidens medicinska kunskap, som uppenbarligen var så hemlig eller kontroversiell att texten behövde krypteras. För Beáta Megyesi började den kryptologiska banan med att hon av en slump drogs in i en chiffergåta. När den amerikanske språkteknologen Kevin Knight höll en föreläsning i Uppsala 2011 efterlyste han ett längre, knivigt chiffer att testa sin datormodell på. Han ville se om hans metod för automatisk översättning även fungerade för krypterad text. Språkforskaren Christiane Schaefer, vid Uppsala universitet, råkade ha fotostatkopior hemma av ett olöst chiffer, som hon hade fått med sig från Tyskland när hon flyttade till Sverige. Ett samarbete inleddes mellan Knight och Schaefer, men det gick inte helt friktionsfritt.

– Hon var en äkta humanist och han var äkta datavetare, så de hade lite svårt att förstå varandra. Jag blev involverad som ett slags mellanhand, säger Beáta Megyesi.

Dokumentet var på 105 sidor och innehöll omkring 75 000 tecken, dels grekiska och latinska bokstäver, dels en mängd symboler lånade från alkemi och astrologi. Det enda som var okodat i dokumentet var ”Philipp 1866” och ”Copiales 3”.

En så kallad Colossus Mark 2-maskin. Bilden är tagen 1943 och Dorothy Du Boisson (till vänster) och Elsie Booker (till höger) var två av de tusentals personer som arbetade på Bletchley Park med att knäcka tyskarnas kod. Majoriteten av de anställda var kvinnor.

THE NATIONAL ARCHIVES UK

INGMARIE ANDERSSON/STOCKHOLMS UNIVERSITET

Beáta Megyesi studerar ett av de många krypterade dokument som finns i den Oxenstiernska samlingen på Riksarkivet i Stockholm.

Tillsammans lyckades de tre forskarna knäcka koden och läsa boken, vars första del var en instruktion till hur nya lärlingar skulle tas upp i en hemlig orden. Texten berättar om hattar som ska tas på, hur hälsningar ska gå till och att mästaren ska plocka ett hårstrå från lärlingens ögonbryn vid ett visst tillfälle under ritualen. Kirurgiska instrument, en magisk amulett och någon sorts avbildning av en trappa finns med. Sedan går texten över till att berätta om frimureri och att kritisera Frimurarorden. För att begripa sammanhanget behövdes även filologisk och religionshistorisk kompetens. Manuskriptet daterades till 1760–1780-tal, men själva texten bedömdes vara ett par decennier äldre, så boken kan vara en avskrift av en annan bok. Den härrör från ett hemligt sällskap som kallas Okulistorden – ögon och seende tycks ha varit centralt. Idéhistorikern Andreas Önnerfors, som forskat om hemliga sällskap i Tyskland på 1700-talet, har beskrivit Copiale-manuskriptet som en rysk docka med olika lager av berättelser. Under lagret av ritualer finns ett mer politiskt budskap om den naturliga rätten till frihet, försvar för upplysningens ideal och kritik mot maktmissbruk. Bara för att en text är dechiffrerad betyder det inte att den blir till fullo begriplig.

MED COPIALE-CHIFFRET TOG Beáta Megyesis forskning en annan riktning. Att de knäckte koden blev en världsnyhet och hon och de andra forskarna började kontaktas av allmänhet och forskare som ville ha hjälp att avkoda krypterad text.

– Jag upptäckte att det inte fanns någon systematisk forskning om historiska chiffer, så jag tänkte att man skulle kunna ta ett större grepp eftersom det finns så många chiffer världen över.

Det började som ett mindre projekt, fortsatte i ett större internationellt forskningsprojekt och nu finns alltså en databas och AI-resurser på plats. Algoritmerna justeras i takt med att databasen fylls på med fler chiffer och nycklar.

I somras träffades de historiska kryptologerna på Bletchley Park och under konferensen Histocrypt, som alltså startades av Beáta Megyesi. Forskarna fick en historisk guidning på godset där den brittiska kodknäckarorganisationen byggdes upp. Trots att hon kunde berättelsen blev Beáta Megyesi tagen av att vara på plats.

– Det var fascinerande att se hur de på ganska kort tid, i all hemlighet, samlade de bästa forskarna för att lösa komplexa problem. Och lyckades! Vilken effektivitet och dedikation! På slutet var det över 9 000 personer som jobbade där och majoriteten av dem var kvinnor.

De flesta bidragen på Histocrypt handlade om avkodning av olika historiska chiffer, med visst fokus på kryptering under andra världskriget. Men en artikel sticker ut – den börjar med en fråga: ”Hur motiverar man en grupp motsträviga tonåringar att lära sig abstrakt matematik en regnig fredagseftermiddag i januari?” Svaret som ges är att historisk kryptologi är ett effektivt sätt att väcka elevernas intresse. Nästan alla har sett filmer eller tv-serier om Bletchley Park, och även om den historiska korrektheten är tveksam så har dramatiseringarna planterat insikten om att kryptering och dekryptering är viktigt, i vissa fall avgörande för historiens utgång. Forskarna medger att det finns svårigheter – mycket av den matematikkunskap som behövs vid dekryptering ingår inte i gymnasieundervisningen. Likväl tror de på chiffrens förmåga att väcka matematisk kreativitet hos tonåringarna. Beáta Megyesi ser ytterligare ett argument för att sätta historisk kryptologi på schemat och det utgår ifrån hennes egen inkörsport till ämnet.

– Jag är starkt övertygad om att universitetsundervisningen har en alltför snäv inriktning i dag i Sverige. Naturvetare har inte de humanistiska perspektiven och humanister är rädda för siffror. Historisk kryptologi har både en algoritmisk, matematisk del, och en humanistisk del i tolkningen av innehållet. Att förlita sig på endast en av dessa aspekter är som att stå på ett enda ben – det ger ingen stabil grund, säger Beáta Megyesi. ●

Vill du lösa historiska chiffer?

Via projektet Decrypts hemsida, de-crypt.org, kan man gå vidare och söka i databasen Decode. Där finns över 8 000 poster och flera hundra av chiffren är fortfarande olösta, så det är fritt fram för vem som helst att försöka dekryptera dem. Många av bilderna är dock tyvärr inte tillgängliga för allmänheten.