Telegrafen på distans

Ett digitalt metodexperiment

Johan Jarlbrink

Ny teknik är inte alltid att lita på. I oktober år 1847 rapporterade *Aftonbladet* att blixten slagit ned i telegrafledningen vid Saint-Germain utanför Paris. Telegrafisten vid den lokala stationen upptäckte ”att telegrafen skref åtskilliga bokstäfver; men som deras sammanhang icke kunde utletas, ärnade han gifva det i sådana fall brukliga tecknet för frasen ’Jag förstår icke’, då han af apparaten fick ett starkt elektriskt slag, åtföljdt af en knall, likasom af ett groft pistolskott”.[1] År 2013 digitaliserades det tidningsnummer som berättar om denna händelse. Tidningen tillhör dem som nu finns sökbara online (på tidningar.kb.se). För att bli sökbara måste de tryckta texterna omvandlas till kod som kan läsas maskinellt. Och med äldre tidningstexter blir det ofta fel. Just den notis som redogör för telegrafins brister år 1847 innehåller i digital form flera av de problem som präglar digitaliserade tidningar. I den textfil som sökbarheten förutsätter heter det att telegrafisten

> såg ett bländande sken långsåt de utmed väggen fastade melalltrådarne som tjena elektriciteten till de visI devärdigavid värdigavid dejemte fullkomen ihåförvintparkerslagna förvintparkerslagna parkerslagna ken —tas till70 70 misvårt fruktarsnart tAf eoch sisrans njes ej ledning och hvilka i detsamma folio ned i stycken ®u ’a så heta att de brände märken i bord och golf[2]

Förutom att en del bokstäver har feltolkats av OCR-tekniken har delar av ord från textspalten intill råkat bli en del av telegraftexten. I denna andra tidningstext handlar det om det tidiga vintervädret i Sankt Petersburg (med förvinter som slagit hårt mot parkerna) och den fruktade koleran (som härjade 70 mil österut). Precis som 1847 har den nya tekniken alltså genererat åtskilliga bokstäver utan sammanhang. Den som använder sig av de digitaliserade tidningsdatabaserna lär inte få någon elektrisk stöt, men riskerar att inte hitta vissa texter eftersom orden har förvrängts och texterna brutits sönder. Syftet i denna artikel är att belysa de problem och möjligheter som digitaliserat källmaterial för med sig. För att tydliggöra såväl bristerna som potentialen kommer jag att utgå från en specifik problemställning: Vad skrivs om den elektriska telegrafen i *Aftonbladet* under 1830-, 1840- och 1850-talen? I vilka språkliga sammanhang placeras den elektriska telegrafen in

i tidningen? Vilka nya möjligheter att undersöka detta har digitaliseringen medfört? Och vilka svårigheter möter i det digitaliserade materialet?

I två inledande avsnitt beskrivs hur det går till när biblioteken omvandlar papperstidningar till digitala filer, samt de ordmassor som processen genererar. Därefter ges en översikt av hur tidigare historisk forskning har tagit sig an liknande textsamlingar med hjälp av digitala metoder, med särskilt fokus på den så kallade distansläsning som litteraturvetaren Franco Moretti utvecklat. I ett följande avsnitt beskrivs hur detta övergripande metodologiska perspektiv omsatts i analysen av *Aftonbladets* telegraftexter. Härefter presenteras resultatet av den digitala textanalysen, hur det kan tolkas och vilka nya insikter det ger, men också de brister som det är behäftat med.[3]

Från papperstidningar till digitala textfiler

På arkiv och bibliotek runt om i världen pågår arbetet med att bevara och tillgängliggöra det pappersbaserade kulturarvet i digital form. Papperet håller i många fall på att vittra sönder och om inget görs kommer viktiga delar av de historiska samlingarna snart att vara borta. Detta gäller inte minst tidningar, som ofta tryckts på sämre papper. Redan mikrofilmningen innebar att tidningarna överfördes till ett medium med längre hållbarhet. Men också mikrofilmsrullar slits, och kvaliteten på de filmade tidningarna var inte alltid den bästa. I digital form återges färger och grafiska element mer verklighetstroget. Dessutom blir texterna sökbara online, så långt tekniken och upphovsrätten tillåter.[4]

Tidningsdigitalisering genomförs i storskaliga projekt med automatiserade processer som redovisar sina resultat i hur många miljoner tidningssidor som gjorts digitalt tillgängliga. Materialet läggs ofta ut i fulltext online i databaser som är sökbara. Vad som tidigare kunde kräva månader av läsning och mikrofilmsrullande kan i bästa fall utföras på någon minut. De digitala databaserna är följaktligen mycket använda och har också påverkat vad historiker forskar om. Den kanadensiske historikern Ian Milligan (2013) har undersökt tidningsreferenserna i samtliga doktorsavhandlingar om Kanadas historia mellan 1998 och 2010. Referenserna till ett antal tidningar ökade med nästan 1 000 procent efter att de digitaliserats i mitten av 2000-talet. Antalet referenser till tidningar som inte digitaliserats ökade också, men bara med några tiotals procent. Eftersom det var några få storstadstidningar som digitaliserades först innebar tillgängligheten att dessa titlar refererades även av dem som undersökte småstäder och glesbygd i andra delar av landet. Ett historiskt storstadsperspektiv riskerade därmed att reproduceras till följd av den nya tekniken.[5]

Eftersom digitaliserat källmaterial får stort genomslag i forskningen är

det viktigt att användarna får insyn i hur själva digitaliseringsprocessen omvandlat det historiska materialet. Att beskriva processen i termer av överföring blir lätt missvisande. Snarare bör den förstås som en form av omvandling eller översättning.[6] Beroende på syftet och intresset hos den som anlitar en källa kan många olika typer av spår hos dokument vara av forskningsmässigt värde. Men det är bara en mindre del av dessa spår som kan reproduceras digitalt. Materiella egenskaper som papperets tjocklek, volymens vikt, hur man bläddrar mellan sidorna och hur de känns i handen är svåra att översätta, och dessa egenskaper prioriteras heller inte av dem som beställer och utför digitaliseringen. Snarare bestämmer man på förhand vilken typ av ”information” dokumenten innehåller, för att sedan koncentrera sig på den. För böcker och tidningar ses ofta den tryckta texten som dokumentets information, men även olika typer av bilder kan prioriteras.[7]

Digitaliseringen utförs i ett antal steg. Bibliotekarier eller arkivarier väljer ut vad och vilka exemplar som ska digitaliseras. Böcker och tidningar skärs ofta upp i ryggen för att skanningen ska gå snabbare. Skannern omvandlar pappersdokumentet till digital kod som sedan kan processas med andra programvaror. Metadata hämtas från externa källor såsom bibliotekskataloger och arkivförteckningar. Ett program för autosegmentering används för att identifiera vilka textsegment på en sida som hör ihop. Ett program för optisk teckenigenkänning tolkar de grafiska tecknen inom varje segment, jämför dem med ett referensmaterial och väljer den bokstav som har störst likhet med det inlästa tecknet, eventuellt med hjälp av en ordbok. Om det fungerar som det är tänkt överensstämmer texterna på den tryckta sidan med de sökbara i databasen.[8] Men resultatet är nästan aldrig så bra.

Notisen om åsknedslaget 1847 illustrerar de vanligaste problemen. 1800-talets tidningar är tättryckta med flera spalter på varje sida. En människa som läser ser direkt att ett tunt streck skiljer den ena spalten från den andra och att texten i den första inte hänger ihop med texten i den andra. Men en maskin har svårare att se detta. I det aktuella fallet har segmenteringsverktyget tolkat det som att textraderna fortsätter från en spalt och ett stycke in i nästa. Resultatet är en helt ny text där åskan i Saint-Germain beskrivs i samma mening som parkerna i Sankt Petersburg. Enskilda bokstäver har också förvanskats. Det är svårt för programvaran att skilja o, e och c, samt l och i, h och li, r och n et cetera. Detta innebär att bokhistorikern som söker på ”folio” besviket får konstatera att det är ett feltolkat ”föllo” som förorsakar felträffen.

Digitala gränssnitt, liksom många andra tekniker, antas ofta vara användarvänliga om användarna inte behöver bry sig om hur den underliggande tekniken fungerar.[9] Också tidningsdatabaserna online präglas av detta tänkande, däribland Kungliga bibliotekets. Användarna tar del av

tidningarna i form av bilder på tidningssidor. Bildkvaliteten är i allmänhet mycket god, och zoomar man in kan man till och med få syn på papperets fibrer. Den som är van att läsa tidningarna i bibliotekens specialläsesalar känner igen sig, även om sidorna i databasen visas en och en och inte som uppslag som i pappersversionerna. Det som gör det möjligt att söka och hitta är emellertid inte dessa fina bilder, utan texterna som ligger dolda bakom bilderna. ”Den teckentolkade texten ska lagras i ett lager bakom bilden som föreställer tidningssidan”, heter det i bibliotekets kravspecifikation.[10] Och här är det inte lika välordnat som det ser ut att vara om man bara ser på bilderna. Den digitala tidningen ser ut att reproducera papperstidningen eftersom de båda liknar varandra visuellt. De digitala texterna skiljer sig dock betydligt från dem som en gång trycktes på papper. Den användarvänlighet som döljer komplikationerna står här i konflikt med källkritikens krav på en medveten materialhantering.

194 miljoner ord

För att få bättre insikt i hur de digitaliserade texterna ser ut kan man för vissa årgångar komma åt själva textfilerna (på data.kb.se). De utgör inga läsvänliga textdokument, utan XML-filer som redovisar varje ord var för sig tillsammans med metadata om vilken rad och vilket textblock ordet ingår i, var på sidan det finns och hur säker OCR-läsningen bedöms vara. Filerna kan emellertid göras om till texter som både kan läsas av människor och processas av programvaror. *Aftonbladet* 1830–1862 omfattar sammanlagt omkring 10 000 fyrsidiga tidningsnummer (i enstaka fall är tidningarna åtta respektive tolv sidor). Sammanlagt innehåller filerna 194 miljoner ord. Antalet unika ord uppgår till nästan nio miljoner. Detta kan jämföras med Andres Fredrik Dalins *Ordbok öfver svenska språket* från 1850-talet med 63 000 ingångar.[11] Men jämförelsen är orättvis eftersom fler böjningsformer, personnamn och geografiska namn, liksom många termer, tillkommer i tidningstexterna. Ändå förefaller antalet orimligt stort. Mängden förvrängda ord är i själva verket gigantisk.[12] Tusentals ord inleds inte ens med en bokstav utan med tecken som !, @ och &. I många fall förefaller bokstäverna att vara korrekta, men orden kapade på mitten: ningar, tion, jas, Stockho. Vissa vanliga OCR-fel kan få stort genomslag. Ordet ”och” är mycket vanligt, men det är även oeh, ocli, ech, osh et cetera. > och < är intressanta specialfall. Detta är vanliga html-koder som används för att ange > respektive < när man designar websidor och andra digitala gränssnitt. Om de verkligen har fyllt någon funktion under digitaliseringsprocessen borde de ha tagits bort innan texterna gjorts tillgängliga för användarna. Den som söker på > i Kungliga bibliotekets tidningsdatabas hösten 2017

får emellertid fler än fem och en halv miljon träffar. Digitala koder är nu en del av det historiska källmaterialet.

För att få en uppfattning om den ordrikedom som OCR-motorn genererar utgick vi från orden ”elektrisk” och ”telegraf” och sökte igenom textmassan med hjälp av ett så kallat Levenshteinavstånd om två bokstäver. Detta innebär att två bokstäver i respektive ord tillåts vara borttagna, tillagda eller utbytta. Med hjälp av denna metod identifierade vi omkring 2 000 varianter av ordet telegraf och 600 varianter av ordet elektrisk. Den högre siffran för ordet telegraf förklaras till stor del av att det förekommer i ett flertal sammansättningar, exempelvis -meddelande, -assistent och -kabel. Varianterna av ordet ”telegraf”, med ändelser och sammansättningar borträknade, är samtidigt många. Vissa felläsningar är vanligare än andra, som att t tolkats som l, e som c och l som t. Variationsrikedomen är dock påtaglig: tele{raf, tele«craf, telbgral, j7elegraf, ulegraf, fesegraf – och hundratals andra kombinationer.

De fritt tillgängliga textfilerna med digitaliserat tidningsmaterial är alltså kvalitetsmässigt bristfälliga på flera sätt, men denna typ av textmassor kan samtidigt erbjuda nya forskningsmöjligheter. Via bibliotekens gränssnitt kan man söka och i bästa fall hitta. Men de bygger samtidigt på att texterna ska läsas en och en, precis som i ett traditionellt tidningsarkiv. Digitala material kan emellertid utforskas även på andra sätt – som dessutom gör en del av kvalitetsproblemen hanterbara.

Digital historia och distansläsning

Att använda datorn till hjälp för att hitta mönster i texter och andra datasamlingar är knappast något sentida påfund. Teologiprofessorn Roberto Busas analys av Thomas av Aquinos samlade verk 1949, stansade ord för ord på hålkort och maskinbearbetade i samarbete med IBM, brukar räknas som det tidigaste exemplet inom humanistisk forskning.[13] Indexeringen och databearbetningen av hundratusentals tidningstexter inom ramen för projektet ”Sverige under andra världskriget” i början av 1970-talet är ett tidigt svenskt försök.[14] Men det var framför allt inom den socialhistoriska forskningen var man snabb med att ta datorn till hjälp. Att den datorunderstödda forskningen inte fick bredare genomslag kan delvis förklaras av den språkliga vändningen (och senare vändningar under 1980- och 1990-tal) och de nya forskningsintressen som följde i dess spår. Användningen av datorer inom den historiska forskningen hade fortsatt betydelse, men främst inom ett antal specifika forskningsgrenar, däribland ekonomisk historia och socialhistoria.[15]

Under de två senaste decennierna har den historiska forskningen, liksom de flesta andra verksamheter, emellertid blivit alltmer beroende av digitala

hjälpmedel. Som bland andra de digitala mediehistorikerna Andreas Fickers och James Mussell påpekat innebär användningen av digitaliserade arkiv och databaser att forskningen i en grundläggande mening blir digital, även om det digitalt mediespecifikas implikationer sällan tydliggörs när forskningsresultaten redovisas.[16] Medan en del ofta kanoniska källor blivit omsorgsfullt digitaliserade i text- och källkritiska projekt blir det stora flertalet källor snarast tillgängliga genom storskaliga digitaliseringsprojekt där den mänskliga kontrollen och kvalitetsgranskningen är minimal. Som informatikern Carl Lagoze framhållit fungerade traditionella arkiv och bibliotek som en kontrollzon där proveniens och autenticitet eftersträvades och garanterades så långt det var möjligt. När samlingarna digitaliseras är kontrollzonen svår att upprätthålla.[17] Inget nationalbibliotek garanterar exempelvis att de miljontals tidningstexter det tillgängliggör digitalt verkligen överensstämmer med dem som en gång trycktes på papper. Här blir det istället den enskilda forskarens ansvar att hantera källkritiken i ett digitalt sammanhang, vilket inte alltid är det lättaste eftersom det är svårt att få insyn i de system som omvandlat samlingarna till digital data. Att endast forska på digitala samlingar av god kvalitet är sällan ett alternativ. Ofta reproducerar digitaliseringsinsatserna äldre materialhierarkier, där det kanoniska (som skönlitteratur och föreställt viktiga nyckeltexter) ägnas textkritisk omsorg, medan digitaliseringen av exempelvis det masspridda och vardagliga tillåts bli som den blir. Många forskningsfrågor kräver emellertid att historiker bearbetar också det svårhanterliga.

Mussell benämner den forskning som främst förlitar sig på digitala databasers sökgränssnitt digital historia 1.0. Digital historia 2.0 utnyttjar i högre grad det faktum att källorna omvandlats till digital data och därmed kan bearbetas maskinellt. Att manipulera och modellera data, att koppla samman data från flera olika källor och att undersöka relationerna mellan data är exempel på detta.[18] När historikern Stephen Robertson utifrån en amerikansk horisont sammanfattar skillnaden mellan det övergripande (eller angränsande) fältet för digital humaniora å ena sidan och digital historia å den andra menar han att det förra domineras av litteraturvetenskaplig textanalys medan det senare ofta använt digitala metoder för att analysera och visualisera data geografiskt (under rubriken ”*Mapping*”).[19] Det är lätt att hitta historiska studier som bekräftar det geografiska intresset, men det är lika lätt att hitta historisk forskning som använder textanalytiska metoder.[20] Och i flera fall säger den disciplinära tillhörigheten mycket lite om de faktiska projektens inriktning. När Ryan Cordell, verksam inom ämnet engelska, studerar de miljontals texter som kopierades från tidning till tidning under det amerikanska 1800-talet är det inte minst de politiska, religiösa, vetenskapliga och ekonomiska idéer som texterna gav spridning

som intresserar.[21] Och som historikern Cameron Blevins forskning om Hustonpressens föreställda geografi kring år 1900 visar kan en geohistorisk analys dessutom med fördel kombineras med en digital textanalys.[22] Finns det någon typ av data i digital form (text, bild, ljud, siffror, metadata) kan den sorteras, kombineras, filtreras och visualiseras med datorns hjälp oberoende av om ingången är idéhistorisk, mediehistorisk, miljöhistorisk eller rättshistorisk. Syftet är ofta att systematisera mycket data på ett smidigt sätt, med det kan också vara att visualisera mönster i mindre material, exempelvis en enskild aktörs resemönster eller sociala nätverk.[23]

Mycket av den historiska forskning som analyserat textsamlingar med digitala metoder tar sitt avstamp i litteraturvetaren Franco Morettis begrepp distansläsning.[24] Distansläsning kan ses som ett metodologiskt perspektiv som innefattar flera olika angreppssätt, exempelvis geografisk visualisering och kvantifiering av metadata.[25] Vanligen innebär metoden att forskaren med datorns hjälp reducerar mångfalden av texter, meningar och ord genom att framhäva det mest (eller minst) frekventa, det ofta samförekommande, hur ordkluster relaterar till varandra och hur frekvenser och ords samförekomster förändras över tid. Antagandet som metoden bygger på är att frekvenser och samförekomster kan säga något om texternas sätt att beskriva världen, vad som är centralt och mindre centralt och i vilka sammanhang som enskilda ord, aktörer, platser och annat placeras i.[26]

Brett definierat, som ”bestämda sätt att tala om världen”, kan distansläsningen förstås som en metod för att kartlägga diskurser på kvantitativ väg.[27] I en studie av Världsbankens språk, så som det framträder i årliga rapporter 1950–2010, analyserar Moretti framväxten av en managementdiskurs genom att undersöka ordfrekvenser, samförekomster och grammatiska förändringar. Rapporterna går från att redogöra för konkreta aktörer som utför eller borde utföra konkreta handlingar (bygga vägar och broar, med cement), till att beskriva abstrakta processer som tycks sakna aktörer.[28] Merparten av sina studier har Moretti emellertid ägnat den engelskspråkiga 1800-talslitteraturen. Den bärande tanken här har varit att låta analysen omfatta mer än den kanon av hundratalet verk som traditionellt intresserat litteraturvetenskapen. Inspirationen har han dock hämtat från Annaleskolan och andra historiker som framhävt de långa perspektiven snarare än enskilda händelser. I inledningen till *Graphs, Maps, Trees* (2005) utgår han från Krzysztof Pomians uppmaning till historiker att skifta fokus ”from the extraordinary to the everyday, from exceptional events to the large mass of facts”.[29] I samma bok argumenterar Moretti även för att den digitala textanalysen skulle vara mer vetenskaplig än den kvalitativa närläsningen. Distansläsningen blir här ett sätt att hantera *mycket* data – och ju mer data desto säkrare slutsatser.[30]

När andra forskare, och inte minst historiker, arbetat vidare med Morettis metoder har de emellertid ofta betonat att distansläsningen bör kombineras med kvalitativ närläsning av enskilda texter. De digitala historikerna Tim Hitchcock och William J. Turkel skriver att distansläsningen av närmare 200 000 rättegångsprotokoll från perioden 1674–1913 låter dem urskilja statistiska mönster som antyder hur rättsprocesserna förändrats över tid, men att denna makroskopiska översikt måste kompletteras med traditionellt arkivarbete för att rättsprocedurerna ska förstås i detalj.[31] Blevins skriver på samma tema:

> Distant reading cannot and will not replace the close reading of historical texts and the interpretation of their meaning and context. The two must be used in concert. Digital methodology played an indispensable role throughout my project in crafting new questions and uncovering hidden patterns. But computational analysis alone cannot articulate the full complexity and significance of my topic. Consequently, I weave together both digital and traditional readings of the text to present a richer understanding of how newspapers produced space.[32]

Distansläsningen kan osynliggöra kontexter, inte minst för att digitala samlingar ofta brister när det gäller att göra kontexterna rättvisa (framför allt publiceringskontexten), men metoden kan också synliggöra kontexter som den kvalitativa fallstudien missar.[33] Distans- och närläsning kan i många fall kombineras med varandra, och forskaren växla mellan de olika perspektiven. Eftersom digitala metoders styrka är att de framhäver övergripande mönster kan de användas av den som vill hitta typiska eller icke-typiska fall att undersöka kvalitativt. Närläsningen kan generera hypoteser som sedan kan prövas på ett större material med digitala metoder – eller omvänt. Eftersom programvara arbetar med hög hastighet lämpar den sig väl för olika typer av experiment. Material kan sorteras och filtreras på nya sätt, nätverken klustras med större eller mindre upplösning, ordklasser och stoppord läggas till eller dras ifrån. Detta antyder också att resultatet av den specifika datakörningen alltid också är resultatet av de instruktioner man ger programvaran. Ett sätt att testa resultatens tillförlitlighet kan vara att upprepa experimentet men ändra några centrala parametrar. Om snarlika mönster uppträder gång på gång är man kanske något på spåren, i annat fall kan mönstren vara resultatet av tillfälligheter.[34]

I Morettis tappning förlitar sig distansläsningen på programvara för att kvantifiera ord och relationen mellan ord. Vad orden betyder och hur resultatet ska tolkas kan inte programvaran ge svar på. Litteraturforskarna Ryan Heuser och Long Le-Khac använder begreppen *signal* och *concept* för

att tydliggöra rollfördelningen mellan maskin och människa. Att exempelvis ett visst antal ord ofta förekommer tillsammans är den signal som programvaran kan generera – men vad detta innebär, vad samförekomsten är ett tecken på, måste en människa sätta ord på. Och här är historisk kunskap och teoretisk förståelse – kompletterad med närläsning – lika fundamentalt som tidigare.[35] I en skrift från hösten 2017, om kopplingen mellan ord som beskriver känslor och ord som beskriver rumslighet i ett stort antal 1800-talsromaner, är det också vad Moretti själv förordar: ”Algorithms generate new facts, whose interpretation continues however to rely on a different hermeneutic tradition.”[36]

Samförekomstanalys

Kan texterna om den elektriska telegrafen i 1800-talets *Aftonbladet* studeras med liknande metoder? För att testa detta måste texterna först hittas och bearbetas. Den felsökning som visar alla de sätt som orden ”elektrisk” och ”telegraf” kan skrivas på åskådliggör inte bara kvalitetsproblemens omfattning, den bidrar även till att göra en del av problemen praktiskt hanterbara. De 2 000 varianterna av telegraf (inklusive ändelser och sammansättningar) och de 600 tolkningarna av elektrisk användes som sökord för att hitta textblock som behandlar den elektriska telegrafen. I drygt 1 250 textblock förekom orden tillsammans. Dessa plockades ut och bearbetades manuellt. Eftersom ett automatiskt genererat textblock nästan aldrig överensstämmer med en text så som den en gång publicerades i tidningen raderade vi de delar av textblocken som innehöll för syftet irrelevanta textfragment. Ett vanligt problem är att segmenteringsverktyget klumpat ihop alla notiser i en spalt till ett enda textblock, och i sådana fall raderade vi alla notiser utom den som behandlade den elektriska telegrafen. Endast ett fåtal textblock raderades helt och de bestod samtliga av innehållsförteckningar med föregående veckas rubriker. Det textmaterial som blev kvar utgjordes av ledarartiklar och referat av tal och debatter, korta essäer och en stor mängd notiser. I textfilerna från 1830-talet hittades endast fyra texter, från 1840-talet hittades 105 och från 1850-talet 998 stycken. Eftersom de textfiler vi fått tillgång till endast innehöll tre år från 1860-talet beslöt vi att inte ta med detta decennium.[37]

Ett sätt att hitta mönster i denna textmassa är att söka efter samförekomster av ord. Genom att kartlägga ord som samförekommer kan man få syn på hur enskilda ord relaterar till andra, vilka grannar ordet har och hur olika grannskap skiljer sig åt. Om den större textmassan delas upp på tidsperiod kan man studera hur relationerna mellan orden förändras över tid. Har man texter från olika källor kan dessa jämföras med varandra.

Samförekomster kan definieras på olika sätt. Vanligen anges ett fönster inom vilket två ord ska förekomma. Här har vi arbetat med en fönsterstorlek omfattande fem ord, vilket innebär att ett rörligt femordsfönster söker igenom varje textblock och registrerar alla ordpar som förekommer inom dessa fönster. Att fönstret är rörligt innebär att det flyttas fram genom texten ett ord i taget. Vidare beräknas hur många gånger varje ordpar förekommer i relation till hur många gånger varje enskilt ord förekommer totalt. Om två ord alltid förekommer tillsammans (exempelvis San och Francisco) finns ett starkt samband mellan dem. Om två enskilda ord är högfrekventa, men samförekommer relativt sällan, är sambandet dem emellan svagare. En stoppordlista med vanliga småord (och, i, är, på et cetera) användes för att filtrera bort mindre betydelsebärande ord. För att begränsa antalet ordpar har vi även satt en minimigräns: ett ordpar måste förekomma i minst sex fönster för att tas med i resultatet. Att fönstret är rörligt innebär att två ord som står bredvid varandra kommer att förekomma i fler fönster än ord som har längre avstånd till varandra. Detta gör att resultatet viktas till förmån för ord som förekommer nära varandra, utan att för den skull hoppa över par med större avstånd.[38]

Fördelen med denna metod är att bruset inte behöver ha någon avgörande betydelse för resultatet. Samförekommer två ord tillräckligt ofta utgör de ett ordpar även om de förvrängts av OCR-tekniken. Om autosegmenteringen gjort att en text som ingår i korpus har förorenats av ord från en helt annan text krävs att samma ord förekommer frekvent också i andra texter för att de ska korrumpera resultatet. I texten om åskan som slagit ned i telegrafledningen utanför Paris samförekommer ordet ”elektriciteten” med bokstavskombinationen ”devärdigavid”, en sammanslagning av ordfragment från spalten bredvid. Eftersom detta ordpar bara förekommer en enda gång i hela textmassan tas det inte med i resultatet. Vad metoden däremot riskerar att missa är lågfrekventa varianter av högfrekventa ord. Som genomgången av alla varianter av elektrisk och telegraf visade kan OCR-tekniken omtolka orden på många olika sätt. De varianter som bara förekommer någon enstaka gång är svåra att fånga upp. Forskare som förespråkar denna och liknande metoder hävdar gärna att de viktigaste orden förekommer så ofta att de ändå kommer med i resultatet. Så länge man inte vet hur frekventa felaktigheterna är och hur de fördelar sig över textmaterialet är det dock svårt att avgöra vilken betydelse OCR-felen har för resultatet. Dessutom är det inte säkert att det alltid är högfrekventa ord och teman som är de mest intressanta. Marginalfenomenen kan vara minst lika värdefulla för analysen och här kan brusnivåerna få större betydelse för utfallet.[39]

I samförekomstsökningen processades texterna från olika decennier separat, med resultatet att 626 ord sammanlänkade i 860 ordpar hittades

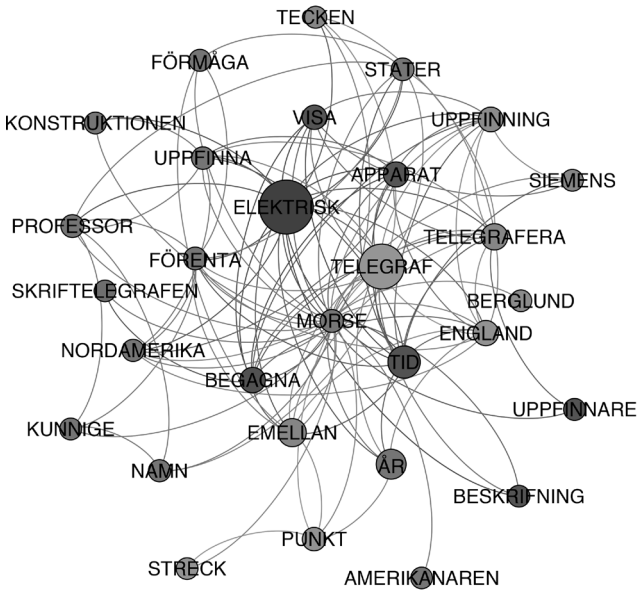
i texterna från 1840-talet och 3 588 ord och 8 964 ordpar från 1850-talet. Ord som förvrängts av OCR-tekniken och som förekommer tillräckligt ofta för att komma med är bland annat ”oeanen”, ”olecriska”, ” Mo«se”, ”Stoeholm” och ”>Stockholms”. För att minska mängden brus slogs sådana ord samman med sina rättstavade motsvarigheter. I enstaka fall har svårtolkade teckenkombinationer bevarats oredigerade. Resultatet har bearbetats i statistik- och visualiseringsprogrammet Gephi, med syftet att identifiera kluster av ofta samförekommande ord.

Den elektriska telegrafen i Aftonbladet

Vad kan vi då säga om *Aftonbladets* telegrafrapportering utifrån de samförekomster som har identifierats? En del av resultaten är högst väntade – vilket i detta fall är bra, eftersom förväntade resultat bidrar till att bekräfta metodens tillförlitlighet. Enligt sociologen och kommunikationsforskaren Patrice Flichy började den populärt hållna tidningspressen att rapportera om elektriska telegrafer omkring år 1837.[40] Den första artikeln som har kommit med i texturvalet är publicerad just detta år.

Den elektriska telegrafens tillkomsthistoria är synnerligen komplicerad, med flera konkurrerande system och successiva förbättringar av flera olika uppfinnare och vetenskapsmän under 1800-talets första halva. Någon enskild uppfinnare går i praktiken inte att urskilja. Flera av de framskjutna efternamnen återkommer bland samförekomsterna: (André-Marie) Ampère, (Louis-François-Clement) Breguet, (Joseph) Henry, (Samuel) Morse, (Charles) Wheatstone.[41] Morse, vars enkla signalsystem fick snabb spridning under 1840- och 1850-talen, är den namngivna person som har flest länkar till andra ord i nätverken, men han förekommer bara enstaka gånger före 1850. De ord som är direkt samförekommande med Morse i 1850-talets *Aftonblad* bildar ett undernätverk vari han framträder som professor och uppfinnare av en ”skriftelegraf”, där kommunikationen sker med hjälp av ”streck” och ”punkter” (figur 1). I *Aftonbladet* är flera personer omnämnda, men efter hand blir telegrafen mer och mer Morses verk.

De första storskaliga experimenten ägde rum på 1830-talet och det var då förslag på telegraflinjer började diskuteras. Samförekomstsökningen fångar upp sträckningarna oavsett om de hunnit realiseras eller ej. De samförekomster som genererats från det fåtal artiklar som publicerades under 1830-talet visar en föreslagen sträckning mellan London och Edinburgh 1837. Under 1840-talet diskuterades och rapporterades om linjer mellan exempelvis London och Paris, Washington och Baltimore, Stockholm och Göteborg. De tidiga experimenten med elektrisk telegrafi utfördes inte minst av tyska och franska vetenskapsmän, men av samförekomstanalysen att döma blev

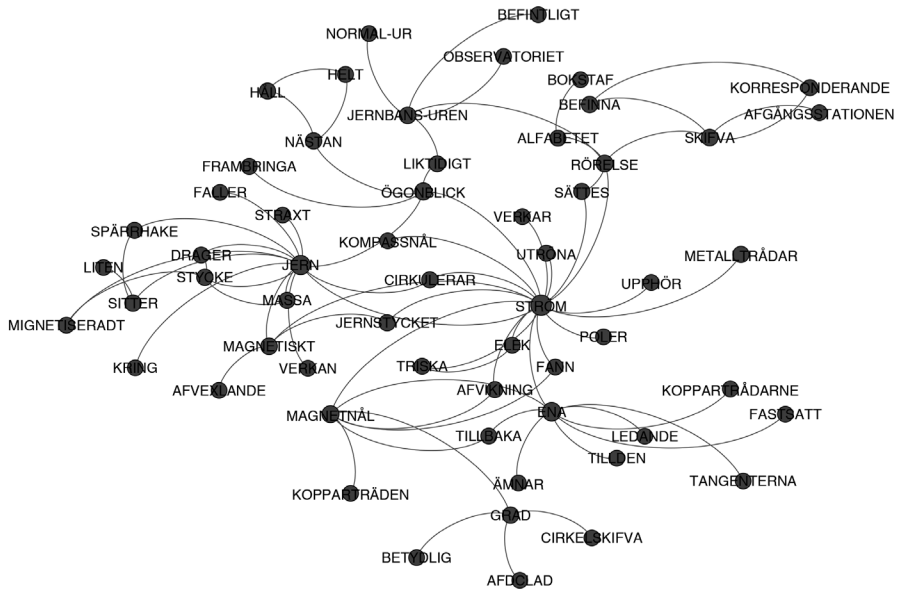


Figur 1: Ord som samförekommer med "Morse" under 1850-talet.

England och London rapporteringens centralnoder under 1840-talet. Under 1850-talet skiftar texternas huvudfokus från den internationella scenen till den inhemska. Nu handlade texterna företrädesvis om teknikens utbyggnad i Sverige.

Telegrafen byggdes inledningsvis ut som ett signalsystem för järnvägen, och järnvägen och synkroniseringen av "jernbans-uren" är också en del av 1840-talets ordnätverk. I övrigt är det i hög grad experiment och konkurrerande tekniker som beskrivs. Texterna berättar om uppfinnare och professorer som arbetar med olika typer av apparater och försök (figur 2). Under 1850-talet finns järnvägen fortfarande kvar, men nu behandlar texterna även telegrafens betydelse för exempelvis börser och tidningspress. Också uppfinnarna finns kvar i texterna, men i den inhemska teknikutbyggnad som beskrivs intar framför allt ingenjörer, regeringen och telegrafverket de centrala positionerna (se nedan).

Dessa iakttagelser bekräftar i hög grad vad som konstaterats i tidigare forskning. En ny metod som endast bekräftar vedertagna synsätt kan förefalla poänglös. Men samförekomstanalysen visar också annat, resultat som avviker i varje fall från telegrafforskningens huvudfåra. För att illustrera distansläsningens poänger och problem kommer jag att uppehålla mig vid tre teman som dominerat forskningen och där samförekomstanalysen ger



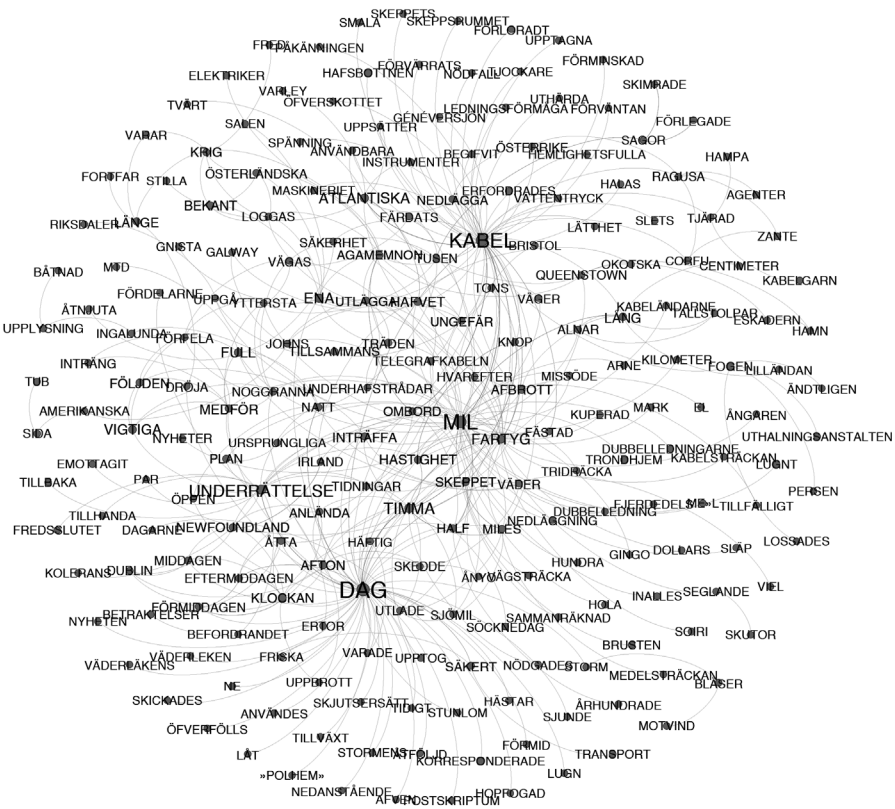
Figur 2: Telegrafens tekniska beståndsdelar i 1840-talets tidningstexter.

delvis andra resultat. Ett första tema gäller hur geografiska förhållanden och avstånd behandlas i texterna, ett andra den telegrafiska kommunikationens föreställda immaterialitet, och ett tredje tema hur den nya kommunikationens utopiska potential diskuteras (och inte diskuteras). Utgångspunkten i det följande är samförekomstanalysens distansperspektiv, men för att illustrera övergripande tematiker lyfts även enskilda textexempel fram.

Avstånd och geografi. Föreställningen om att den elektriska telegrafen gjorde kommunikationen oberoende av geografi och avstånd etablerades tidigt. Rumslig kommunikation hade tidigare varit synonym med fysisk transport och tiden det tog att ta sig mellan två punkter (undantaget signalerna som överfördes med semaforen och andra optiska medier). På ett teoretiskt plan innebar den elektriska telegrafin istället en blixtsnabb signalöverföring mellan sammankopplade stationer (och när tidningstexterna inte rapporterade att blixten slagit ned i ledningarna skrev de gärna att telegrafmeddelanden skickades ”med blixtens hastighet”). Telegrafins uppfinnare och förespråkare bidrog själva till att ge dessa idéer spridning.[42] I senare forskning har framför allt medieforskaren James Careys karakteristik av telegrafen fått genomslag: ”the telegraph freed communication from the constraints of geography”.[43] Mediehistorikern Jonas Harvard har sammanfattat detta huvudspår i forskningen och konstaterat: ”This development is

often described in terms of a 'time-space compression', the 'annihilation of time and space' or similar constructs."[44]

Att telegrafen innebar att kommunikationen i tid och rum förändrades går inte att förneka – alla nya medier har sådana effekter.[45] Däremot är det inte säkert att de människor som främst kom i kontakt med den nya tekniken via tidningstexter upplevde att geografi och avstånd plötsligt blivit överspelade. Exakt angivna avstånd skrevs återkommande ut i *Aftonbladet* varje gång en ny telegraflinje planerades eller togs i bruk. Ordet "mil" hör till de mest centrala (alltså samförekommande med ett flertal andra ord) under såväl 1840- som 1850-talet (se figur 3). Det var sannolikt få läsare som tidigare visste exakt hur långt det var mellan olika orter i Sverige och i utlandet. I samband med att nya telegrafledningar drogs fick de emellertid veta att det var 200 engelska mil mellan Holyhead och London, tolv och en halv svenska mil mellan Baltimore och Washington, 1 647 engelska mil



Figur 3: I 1850-talets tidningstexter om telegrafen är geografiska förhållanden och materiella beståndsdelar återkommande teman.

mellan nordamerikanska Saint John och irländska Galway, 31 mil mellan Christiania och Göteborg et cetera.[46] Vad som tidigare kunde ha framstått som en diffus geografi bestående av avlägsna platser blev med tidningsrapporterna om telegrafens dragningar till något karterat och uppmätt, med avstånden angivna i exakta mil.

Denna upptagenhet vid geografiska förhållanden märks också på andra sätt. Bland de samförekommande orden som genererats finns ett flertal som refererar till de vatten- och landmassor som avgjorde var och hur kablarna kunde dras: ”fastlandet” och ”hafsbotten”, ”djup” och ”grund”, ”sand” och ”klippor”, ”öar” och ”sjöar”, ”uddar” och ”hafsvikar”. Rapporteringen under 1840- och 1850-talen handlade i hög grad om ”anläggandet”, telegraflinjernas ”sträckning” och de fysiska förutsättningar och prövningar som olika platser erbjöd. Under 1850-talet gjordes flera försök – mer eller mindre lyckade – att lägga ut telegraftrådar på havsbottnarna. Atlanten, Svarta havet, Öresund och Engelska kanalen hörde till de vatten där kablar drogs. *Aftonbladet* rapporterade regelbundet och stundtals mycket detaljerat om dessa företag. I samförekomstanalysen av 1850-talets texter kan ett omfattande ordkluster urskiljas som fångar in huvuddragen i dessa projekt och planer (figur 3). Här syns fartyg och kablar, stormar och missöden, och det hårda arbetet med att bygga upp det nätverk som den elektriska kommunikationen förutsatte. I dessa texter framträder inte geografin som en dimension som tekniken gjorde överspelad. Geografins formationer blir stundtals kommunikationens själva förutsättning. Så här kunde det exempelvis låta när en planerad underhavstelegraf mellan Europa och Nordamerika diskuterades 1855:

> ’På hafvets botten, emellan cap Race i Newfoundland och cap Clear i Irland, finnes en märkvärdig stepp, hvilken redan är känd under namn af den telegrafiska platåen. Afståndet mellan dessa två kustlinjer är 1600 mil och utefter hela sträckan är hafvet sannolikt på intet ställe djupare än 10,000 fot.’ [---] Det är liksom en särskild anordning af naturen för att befrämja detta stora företag. En kedja af bergsplatåer ligger der på hafvets botten, liksom den vore enkom gjord för att uppbära den elektriska kedjan längs hafvets djup. [---] Man har funnit att hafsbottnen består icke af skarpa och branta klippor utan af en mjuk sand i hvilken telegrafkabeln kan nedsjunka och inbäddas för sekler.[47]

Den signal som samförekomstsökningen genererar består alltså till stor del av ord som antyder att geografiska förhållanden och avstånd var en viktig del av det sammanhang som den elektriska telegrafen placerades in i. Närläsningen av enskilda texter visar hur de geografiska sammanhangen kunde framhävas som teknikens själva förutsättning. Min tolkning av detta

är att under de decennier då telegrafnätverken byggdes upp var geografiska förhållanden och avstånd inget som suddades ut eller hamnade i bakgrunden. Tvärtom synliggjorde telegraflinjerna geografins betydelse och vad som dolde sig under havsytan. När kablar rullades ut tydliggjordes exakt hur långt det var mellan olika orter. Efter några decennier, när tekniken fanns på plats och fungerade som det var tänkt, är det fullt möjligt att den krympte rummet. Men under de år då teknikutbyggnaden fortfarande var en nyhet framhävdes istället helt nya rum, ibland 10 000 fot under havsytan. Samförekomstanalysen visar att fysiska avstånd och geografiska förhållanden var ett bärande tema i rapporteringen.

(Im)materiell kommunikation. Idén om att telegrafen upphävde geografiska avstånd är intimt förknippad med föreställningen om den elektriska telegrafen som en immateriell kommunikationsform. Telegrafen hade förmågan att krympa rummet just eftersom kommunikationen inte krävde att meddelanden transporterades fysiskt. Även denna tanke etablerades tidigt av teknologins innovatörer och uttolkare. Mediefilosofen John Durham Peters skriver om telegrafen i sin idéhistoriska genomgång av kommunikationsbegreppets omvandlingar att ”its cargo is weightless and swift”. Han menar att tekniken placerades in i en längre tradition av andlig kommunikation: telegrafen ”fits precisely into the lineage of Augustine, the angels, and Mesmer: communication without embodiment, contact achieved by the sharing of spiritual (electrical) fluids”.[48] I senare forskning har den andliga dimensionen tonats ned, men idén om den immateriella kommunikationen har som regel levt vidare. I en forskningsöversikt, publicerad 150 år efter att den första (fungerande) Atlantkabeln invigdes 1866, konstaterar historikern Simone Müller att ”telegraphy seemed to have dematerialized information” och att denna ”loss of materiality in communications” var en viktig förutsättning bland annat för den moderna börshandeln.[49] Jonas Harvard sammanfattar forskningen på följande vis: ”It is commonly argued that the development of the electric telegraph led to the dematerialization of communication”.[50]

Vad har distansläsningen att bidra med vad gäller telegrafins föreställda viktlöshet? För det första kan det konstateras att den ”andliga” signalen inte är särskilt stark när *Aftonblad*-texterna analyseras digitalt. Det som genomsyrar telegrafrapporteringen är snarast den immateriella kommunikationens motsats (eller materiella förutsättning, om man så vill): stolpar, koppartråd, guttaperka, batterier, magneter och apparater. ”Tråd”, ”jern” och ”kabel” hör till de ord som samförekommer med flest andra ord under såväl 1840- som 1850-talet. En del av detta syns i figur 3, där flera av anläggningsprojektens nyckelord finns med: tallstolpar och kabel, tråd och kabelgarn, fogar och ändar, dubbla och tjärade, långa och tjocka. En del av den tekniska vokabulären har även samlats i ett av 1840-talets ordkluster

(figur 2). Här refererar orden mindre till utläggandets komplikationer och mer till uppfinningen och experimenten: järn och koppar, magneter och nålar, spärrhake och cirkelskiva.

En tolkning som ligger nära till hands är att den telegraf som tidningsläsarna fick beskriven för sig var en teknisk och materiell infrastruktur. En närläsning av enskilda texter visar att detaljnivån kunde vara närmast extrem. När en telegraflinje skulle anläggas mellan England och Frankrike år 1851 hade det varit ett bra tillfälle att beskriva de andliga eller kulturella konsekvenserna av att en blixtsnabb förbindelse höll på att upprättas mellan de forna fienderna. Men i *Aftonbladet* valde man istället att beskriva vad som utmärkte själva kabeln:

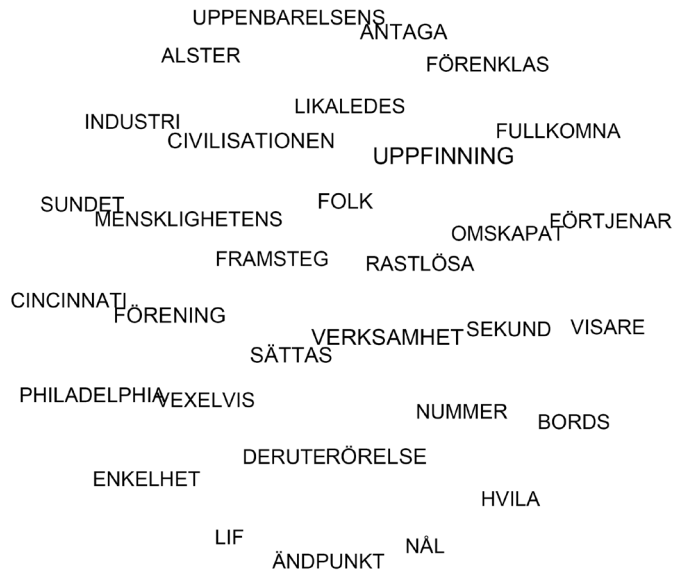
> Dessa trådar äro af koppar, öfverdragna med svavelbunden guttaperka och öfverspundna med i kautsjulösning indränkt hampgarn, samt ytterligare bevarade medelst ett öfverdrag af vattentät stenkolsfernissa. Fem sålunda förberedda trådar bilda perterna i ett kabeltåg, på lika sätt fernissadt. Sjelfva kabeltåget är vidare öfverspunnet som en bas-sträng, men med grof förzinkad jerntråd, och alltsammans ytterligare öfverdraget med en i stenkolstjära indränkt kabelgarnsspånad.[51]

Också vad gäller kommunikationens materialitet kan det förmodas att telegrafen efter några decennier fick en delvis annan betydelse. Kanske sågs den då mer som en snabb kommunikationskanal och mindre som en fysisk infrastruktur (eller inte: telegraftrådarna hängde ofta fullt synliga). I *Aftonbladets* texter från 1840- och 1850-talen är det emellertid en högst materiell kommunikationsform som beskrivs.

Teknikens utopiska potential. Nya tekniker lanseras ofta med ett löfte om de fantastiska möjligheter de bär med sig. Med introduktionen av nya medietekniker följer utopiska idéer, fantasier och drömmar.[52] Den elektriska telegrafen är inget undantag. Möjligheterna till perfekt kommunikation antogs, enligt den utopiska retoriken, kunna leda till civilisatoriska framsteg, ökad samhörighet människor emellan, fred på jorden och så vidare.[53] I sin analys av de praktiska svårigheterna med att få den telegrafiska tekniken att faktiskt fungera drar Jonas Harvard slutsatsen att ”teknikens huvudsakliga betydelse låg på det symboliska planet”. Det kommunikationsnät som under sina första decennier drabbades av ständiga avbrott fick på det retoriska planet representera ”den framtidsriktade modernitetens drömmar om ett ögonblickligt idéutbyte”.[54]

Att dessa föreställningar verkligen existerade är väl belagt, men frågan är hur spridda de var. Återfinns denna retorik i *Aftonbladet* under 1840- och 1850-talen? I samförekomstsökningen i 1840-talets texter hittas endast några

enstaka nyckelord som antyder en utopisk diskurs (”framtid”, ”frihet”), men det går inte att identifiera något enskilt kluster som refererar till sådana texter. För 1850-talet går det däremot att hitta ett litet kluster där en handfull ord ser ut att beskriva teknikens fantastiska möjligheter: ”civilisationen”, ”framsteg”, ”fullkomna”, ”omskapat”, ”menskilghetens” (figur 4). Andra ord som återfinns spridda i helt andra kluster pekar åt samma håll: ”gemenskap”, ”fred”, ”upplysning”, ”vänskap” och ”rättvisa”. ”Civilisationen” och ”framtid” samförekommer med flest andra ord (13 respektive 8), medan de andra endast är länkade till något enstaka ord vardera. I denna distansläsning ser det alltså ut som att det kan finnas tidningstexter som uttrycker utopiska tankar, men att temat samtidigt existerar i marginalen.



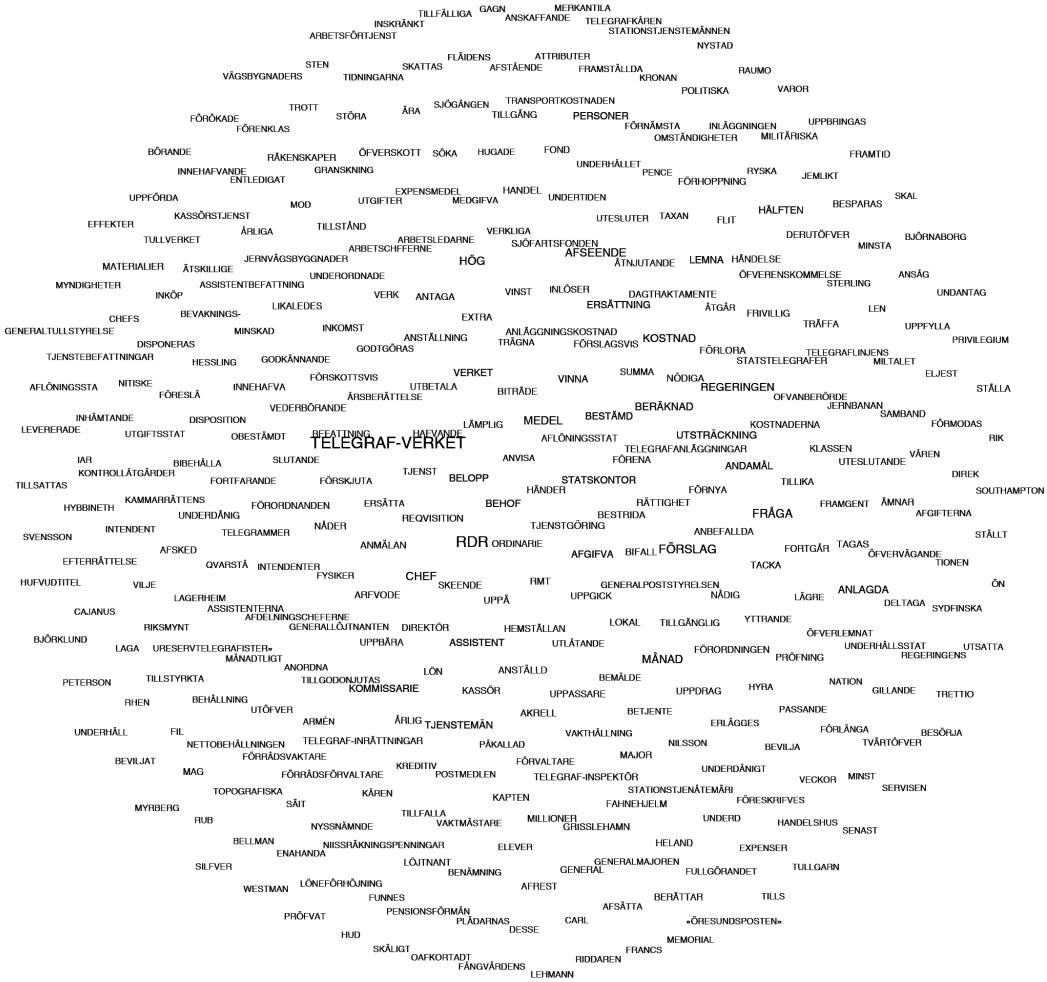
Figur 4: Telegrafins civilisatoriska möjligheter i 1850-talets texter.

En vidare granskning av textmaterialet gör gällande att det bara är några få texter som beskriver telegrafens fantastiska möjligheter, men att de är relativt ordrika – i första hand kortare essäer, ledarartiklar och referat av tal. År 1851 publicerade tidningen exempelvis ett utdrag ur den amerikanska filantropen och fredsaktivisten Elihu Burritts skrift *Ett oliveblad för folket*, under rubriken ”Tankar om krig”. Burrit tyckte sig där se flera tecken på en ljusare framtid och världsfredens förverkligande. Till de tecken på civilisationens framsteg han räknade upp hörde kanalbyggen, järnvägen och den elektriska telegrafen:

> Se der de yttre kännetecknen på verkningarna af idéen om *allmän förbrödring*, hvilken nu bemäktigat sig folkandan i alla länder, samt förbereder nationerna för det tillstånd af endrägt och harmoni som uppenbarelsens ord lofvat menniskoslägtet. Det är civilisationens på de mekaniska och fysiska krafternas tillgodogörande offrade ansträngningar, som äro bestämde att i förkroppsligad form framställa den höga sanningen: 'Gud har danat hela meeniskoslägtet af ett och samma blod!'[55]

I ett antal liknande texter återfinns flera av de utopiska idéernas nyckelord, ofta i samband med att skribenterna tog ett vidare grepp på civilisatoriska framsteg inom kultur och industri.[56] Det är emellertid viktigt att komma ihåg att samförekomstsökningen fångar upp orden även om texten ifråga avvisar teknikens välsignelser. I en refererad debatt om läroverksfrågan 1858 talade en medverkande om behovet av klassisk bildning eftersom "den materiella civilisationens vingar skola icke lyfta henne uppöver telegrafstolparne".[57] Drömmarna om telegrafins möjligheter går att spåra, men relativt sett är det endast en mindre del av texterna som behandlar dessa teman.

Texterna i *Aftonbladet* gällde i huvudsak helt andra saker, och om samförekomstsökningen ibland har problem att fånga upp det marginella är metoden desto mer effektiv när det gäller att synliggöra det högfrekventa. Ett omfattande ordkluster baserat på 1850-talets texter framhäver vad som närmast kan betraktas som den utopiska diskursens motsats: telegrafins byråkrati (figur 5). Här återfinns den personalstyrka som hade till uppgift att hantera telegraferna: inspektörer och intendenter, kassörer och assistenter, kommissarier och vaktmästare, direktörer och majorer. Den värld som här beskrivs består av utlåtanden och bifall, förordnanden och kontrollåtgärder, utgiftsstat och årsberättelser. I texterna om telegrafverkets organisation och arbete handlade det inte om drömmar och utopier, utan om medel och kostnader, befattningar och förslag. En närmare granskning av texterna visar att en omfattande del av tidningens telegrafrapportering består av korta texter som refererar förslag, beslut och ekonomiska förhållanden inom de statliga organ som administrerade telegrafin och dess utbyggnad. Eftersom tidningsredaktionen var liten och hade begränsade resurser var det svårt att fylla hela tidningen med egenskrivna texter. En stor del av innehållet bestod istället av kungörelser, lagtexter, årsberättelser, protokoll och offentlig statistik, med statliga instanser som Telegrafverket som avsändare.[58] Bland *Aftonbladets* telegrafrelaterade texter återfinns refererade riksdagsdebatter och regeringsbeslut, frågor om patent och ekonomiska anslag, regler för hur telegrafen skulle få användas och hur mycket det skulle kosta, tvister om markanvändning och importtullar på materiel, telegrafpersonalens



Figur 5: Telegrafins byråkrati i 1850-talets tidningstexter.

arbetsvillkor och så vidare. Huvuddelen av *Aftonbladets* telegraftexter såg ut ungefär så här:

> *Wenersborg* d. 10 Jan. Kongl. elektriska telegrafverket annonserar om uppköp under innevarande Januari och nästkommande Februari månader af 220 st. tallstolpar, 14 alnar långa och minst 4 tum i lilländan, på hvarje mil härifrån till Uddevalla och vidare till norska gränsen förbi gästgifvaregårdarne Gesäter, Lundby och Högen, för den telegraflinie, som nästa år derstädes kommer att anbringas.[59]

Medieforskaren Brian Winstons har argumenterat för att nya kommunikationsteknologier ofta bär på en radikal potential, men att deras faktiska betydelse som regel begränsas så snart marknadskrafter och statsbyråkratier tar kontrollen över dem.[60] Den metod som använts här kan varken bekräfta eller avfärda detta. Vad som visats är emellertid att byråkratins språkbruk rent kvantitativ övertrumfade de utopiska visionerna. Ordklustret i figur 5 utgör drygt tolv procent av 1850-talets totala ordnätverk. Nätverket i figur 4, där en handfull ord tycks referera till teknikens fantastiska möjligheter, representerar endast en procent av den sammantagna ordmängden. Med distansläsningens metoder framträder telegrafen som något som administrerades, reglerades och budgeterades.

Detta resultat kan illustrera övergripande skillnader mellan olika metodologiska angreppssätt och vilka konsekvenser de får. Mediehistorisk forskning har haft att hantera och orientera sig i mycket omfattande empiriska material, ofta helt utan arkivförteckningar och sökverktyg. Att forskningen i det läget gärna följt redan upptrampade stigar och i första hand uppmärksammat det avvikande, exceptionella och kanoniska är inte förvånande – det har helt enkelt varit svårt att överblicka, hitta och systematisera det till synes oansenliga och vardagliga. Mussell konstaterar att ”there has been an institutionalized preference for the exceptional – what makes a particular text or publication important or different – over the repetitive and generic”. Och en viktig anledning är just de metoder och verktyg som funnits till förfogande: ”without the tools and methodologies to interrogate the repetitive (and so the generic and the abundant), analysis is restricted to generalizations based on the exceptional without really establishing the grounds for exceptionality within the culture of the period”.[61]

Med digitala verktyg är det enklare att hitta mönster i stora textsamlingar, men här kan det istället vara svårt att göra rättvisa åt exceptionella enskildheter. Att den elektriska telegrafen emellanåt förknippades med utopiska idéer är *också* intressant, och säkert inte oviktigt i sin samtid, men i en samförekomstanalys reduceras temat till ett marginalfenomen som knappt ger några avtryck. Tidningstexterna som refererar byråkratins agerande är *verkligen* frekventa, men metoden som använts riskerar måhända att blåsa upp dem mer än vad som kanske är befogat. Det byråkratiska språket, liksom det vetenskapliga, är som gjort för den digitala textanalysen: terminologin är i hög grad standardiserad och samma ord återkommer i text efter text. Att en del av orden förbises på grund av att de förvrängts av OCR-motorn påverkar inte nödvändigtvis den övergripande bilden – de korrekta varianterna är ändå så många. Den utopiska prosan är betydligt mer varierad och icke-standardiserad, därmed blir samförekomsterna inte lika frekventa. Är en del av orden dessutom förvrängda blir tematiken än

svårare att urskilja. Eftersom olika metoder premierar olika resultat är det rimligt att distansläsningen av större textmassor kombineras med en närläsning av enskilda texter.

Avslutning

Databaserna med digitaliserade tidningstexter skapar nya möjligheter att utforska historiska diskurser med hjälp av digitala metoder. De textsamlingar som digitaliseringen genererat är samtidigt svårbearbetade. Att arbeta direkt med de digitala filer som sökgränssnittet döljer kräver en hel del manuella insatser, men innebär också att forskaren kan uppmärksamma och åtgärda en del av de problem som präglar materialet.

Med tanke på materialets beskaffenhet, och att själva digitaliseringsprocessen omvandlat en del av tidningsinnehållet till svårigenkännligt brus, har min distansläsning av telegrafen i tidningen främst tagit fasta på högfrekventa ord och övergripande teman. En del av resultaten bekräftar tidigare forskning. Uppfinnare och ingenjörer som brukar lyftas fram framträder tydligt även när texterna om telegrafen läses på distans. Att tekniken inledningsvis i första hand var en tysk och engelsk angelägenhet och att den senare placerades in i en svensk kontext är ett resultat som heller inte förvånar. I andra avseenden synliggör den digitala textanalysen delvis andra aspekter av telegrafen än dem som dominerat i tidigare forskning. För att vara en teknik som påstås ha frigjort kommunikationen från geografins begränsningar är geografiska förhållanden oväntat närvarande i de texter som beskriver tekniken. På distans framstår det som att anläggandet av nya telegraflinjer synliggjorde den geografi som tekniken skulle ”förinta”. Ett näraliggande tema är den telegrafiska kommunikationens föreställda immaterialitet. Tidigare forskning har i hög grad reproducerat idén om telegrafin som en immateriell kommunikationsform. I den digitala textanalysen framträder tekniken istället som en högst materiell infrastruktur. Den som tog del av *Aftonbladets* texter om telegrafen fick intrycket att det var på kopparkablar och trästolpar kommunikationen vilade. En tredje slutsats som kan dras av denna distansläsning är att de utopiska idéer som telegrafen ofta förknippats med i varje fall inte var särskilt framträdande i den tidning som under perioden var Sveriges största. Betydligt större utbredning hade den byråkratiska diskurs som kan betraktas som de utopiska idéernas motsats. I dessa avseenden kan distansläsningen av många texter alltså fungera som ett komplement till den närläsning av förmodade nyckeltexter som tidigare forskning förlitat sig på.

Att läsa texter på distans är samtidigt att strunta i en del av det som vanligtvis gör texter meningsfulla: Vem har skrivit? Inom vilken genre? Vilket

perspektiv är det som förmedlas? Vad argumenterar texten för? Vad handlar den om mellan raderna? Digitala metoder för textanalys kvantifierar ord och relationen mellan ord, ingen text är nödvändigtvis mer betydelsefull än någon annan (så länge den inte innehåller fler ord). Det som kan göra den enskilda texten orättvisa kan emellertid vara det som gör den större textmassan rättvisa. Det typiska för 1800-talets tidningar är det repetitiva. Varje dag nya texter, men samtidigt snarlika gårdagens och förra årets. Att plocka ut enskilda guldkorn blir lätt missvisande, mer representativa är de periodiskt återkommande alldagligheterna.[62] Där bibliotekens gränssnitt premierar den enskilda sökträffen gör digitala analysmetoder det möjligt att få syn på det återkommande. Ledarartikeln om telegrafins kulturella innebörd är måhända mer intressant än notisen om Telegrafverkets uppköp av trästolpar, men betraktar man textmassan på distans är det svårt att ens urskilja ledarretoriken för alla trästolpar.

Även om den metod som här har använts gör en del av kvalitetsproblemen hanterbara måste resultaten tolkas med försiktighet. Många av de brister materialet uppvisar kan inte avhjälpas av en enskild forskare. *Aftonbladet* var en av de första tidningarna som Kungliga biblioteket skannade in, den programvara som användes var ny och de som ansvarade för processen hade begränsad erfarenhet. Nedslag i de tidningsfiler som genererats några år senare, exempelvis för *Dagens Nyheter*, antyder ett något bättre resultat: andelen ord som tolkats fel av OCR-motorn är något mindre, segmenteringsproblemen inte lika påtagliga. Eftersom tekniken blir bättre och bättre, och de ansvariga efter hand utökar sina kunskaper, finns möjlighet att processa redan inskannat material på nytt, med ett förhoppningsvis bättre resultat. Detta skulle göra databaserna och textfilerna mer användbara, men skulle också innebära att vedertagna föreställningar om arkivet som något fixerat och stabilt ställs på ända. Den digitaliserade tidningen är inte densamma som den som en gång trycktes, men om delar av processen görs om innebär det även att den ena digitala versionen inte är den andra lik: ”The aesthetics of fixed order is being replaced by permanent reconfigurability”.[63] I digital form omskapas det historiska källmaterialet med varje uppdatering.

The telegraph from a distance: An experiment with digital method

The digitization of historical records not only creates new research opportunities but also challenges. Documents may be searched online and digital tools may be used for finding patterns in large datasets. Yet, the quality of digitized material is often low and it may be difficult for researchers

to evaluate this material when hidden behind interfaces. My aim in this article is to explore a digitized newspaper and the associated problems and opportunities. My exploration is guided by the following question: What was written about the electrical telegraph in *Aftonbladet* in the 1830s, 1840s and 1850s?

The files of the digitized *Aftonbladet* for the years 1830–1862 contain massive amounts of corrupted words generated by the OCR engine. Another problem is that the tool used for auto-segmentation has merged separate text items into single text blocks and split long texts into smaller parts. The digitization process is not a neutral text transfer from one medium to another, it is a process generating new texts never printed in the original newspaper.

In order to find as many (corrupted) versions of the words "electrical" and "telegraph" as possible, the text files were searched with a Levhinstein distance of two (allowing for two letters to be added, replaced or missing). The words that were found were used as search words to identify texts blocks reporting on the electrical telegraph. To find themes in the texts, a co-occurrence search was carried out, identifying clusters of words frequently co-occurring. Previous research has focused on utopian ideas associated with this technology and the telegraph as an immaterial form of communication obliterating physical space. The clusters identified in the digital analysis suggest that more mundane topics dominated the newspaper: the electrical telegraph was described as a material form of communication controlled by bureaucracy, depending on and emphasizing geographical conditions. Digital databases make new research possible, but the low quality of digitized texts results in uncertainties concerning the results. Researchers using these databases need to be aware of the digital forms and interfaces regulating research practices.

Keywords: Digital methods, databases, digitized newspapers, media history, telegraph

Noter

1 "Åskan och elektriska telegrafen", *Aftonbladet* 16/10 1947.

2 Den digitaliserade tidningen finns tillgänglig i XML-format, "bib4345612_18471016_0_s_0003_alto.xml", tillgänglig på https://data.kb.se/datasets/2014/10/aftonbladet/1847/10/urn%253Anbn%253Ase%253Akb%253Adark-41615/ (2017-03-30).

3 Studien ingår i projektet "Digitala lägg", finansierat av Torsten Söderbergs stiftelse. Den digitala textbearbetningen är gjord i samarbete med Roger Mähler vid Humlab, Umeå universitet.

4 James Mussel, *The Nineteenth-Century Press in the Digital Age*, Basingstoke 2012.

5 Ian Milligan, "Illusionary order: Online databases, optical character recognition, and Canadian history 1997–2010", *The Canadian Historical Review* 2013:5, s. 541f.

6 Mats Dahlström, ”The complete edition”, i *Text Editing, Print, and the Digital World*, Marilyn Deegan & Kathryn Sutherland (red.), Farnham 2009, s. 31.
7 Marlene Manoff, ”The materiality of digital collections: Theoretical and historical perspectives”, *portal: Libraries and the Academy* 2006:3, s. 313–316.
8 Om digitaliseringsprocesser, se Lars Björk, *How Reproductive is a Reproduction? Digital Transmission of Text Based Documents*, Borås 2015.
9 Jussi Parikka, *What is Media Archaeology?*, Cambridge & Malden 2012, s. 113.
10 ”Kravspecifikation DD v.1.2” (daterad 2011-04-29), http://digidaily.blogg.kb.se/dokument/ (2017-03-30)
11 Se Dalins ordbok i Språkbankens digitala version, https://spraakbanken.gu.se/swe/resurs/dalin (hämtad 2017-03-30).
12 En mer utförlig analys av det digitala bruset presenteras i Johan Jarlbrink & Pelle Snickars, ”Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive”, *Journal of Documentation* 2017.
13 Susan Hockey, ”The history of humanities computing”, i *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens & John Unsworth (red.), Oxford 2004.
14 Johan Jarlbrink, ”Historievetenskapens mediehantering”, i *Massmedieproblem: Mediestudiets formering*, Mats Hyvönen, Pelle Snickars & Per Vesterlund (red.), Lund 2015.
15 För en översikt över den datadrivna historievetenskapens framväxt, se William G. Thomas, ”Computing and the historical imagination”, i *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens & John Unsworth (red.), Oxford 2004.
16 Andreas Fickers, ”Towards a new digital historicism? Doing history in the age of abundance”, *VIEW: Journal of European Television History and Culture*, 2012:1, s. 23f; Jim Mussell, ”Doing and making: History as digital practice”, i *History in the Digital Age*, Tony Weller (red.), London & New York 2013, s. 81.
17 Carl Lagoze, ”Big Data, data integrity, and the fracturing of the control zone”, *Big Data & Society* 2014:2, s. 6f.
18 Mussell 2013, s. 91.
19 Stephen Robertson, ”The differences between digital humanities and digital history”, i *Debates in the Digital Humanities 2016*, Matthew K. Gold & Lauren F. Klein (red.), Minneapolis 2016.
20 En översikt över den geohistoriska forskningen ges i David J. Bodenhammer, ”The spatial humanities: Space, time and place in the new digital age”, i *History in the Digital Age*, Tony Weller (red.), London & New York, 2013.
21 Se projektbeskrivningen för Ryan Cordell & David Smith, ”Viral texts: Mapping networks of reprinting in 19th-century newspapers and magazines”, http://viraltexts.org (2017-10-13).
22 Cameron Blevins, ”Space, nation, and the triumph of region: A view of the world from Houston”, *Journal of American History* 2014:1.
23 Se t.ex. analyserna av Herodotos *Historia* i *New Worlds from Old Texts: Revisiting Ancient Space and Place*, Elton Barker, Stefan Bouzarovski, Christopher Pelling & Leif Isaksen (red.), Oxford 2016.
24 Tim Hitchcock & William J. Turkel, ”The *Old Bailey Proceedings*, 1674–1913: Text mining for evidence of court behavior”, *Law and History Review* 2016:4, s. 934; Andrew Stauffer, ”Introducing: Searching engines, reading machines”, *Victorian Studies* 2011:1, s. 64; Blevins 2014, s. 125f.
25 Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*, London & New York 2005, s. 1f.
26 Franco Moretti, ”Conjectures on world literature”, *New Left Review* 2000:1, 56f.

27 Jämför Paul DiMaggio, Manish Nag & David Blei, ”Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding”, *Poetics* 2013:6, s. 578.
28 Franco Moretti & Dominique Pestre, ”Bankspeak: The Language of World Bank Reports”, *New Left Review* 2015:92.
29 Krzysztof Pomian citerad i Moretti 2005, s. 3.
30 Moretti 2005, s. 2–4.
31 Hitchcock & Turkel 2016, s. 955.
32 Blevins 2014, s. 126.
33 Fickers 2012, s. 22.
34 En beskrivning av den digitala textanalysen som en experimentell process presenteras i Sarah Allison, Ryan Heuser, Matthew Jockers, Franco Moretti & Michael Witmore, *Quantitative Formalism: An Experiment*, Pamphlets of the Stanford Literary Lab, Stanford 2011. Tillgänglig på https://litlab.stanford.edu/pamphlets/ (2017-10-13). Se även Bernhard Rieder & Theo Röhle, ”Digital methods: Five challenges”, i *Understanding Digital Humanities*, David M. Berry (red.), Houndmills, Basingstoke & Hampshire 2012, s. 77–79.
35 Ryan Heuser & Long Le-Khac, ”Learning to read data: Bringing out the humanistic in the digital humanities”, *Victorian Studies* 2011:1, s. 81f.
36 Franco Moretti, *Patterns and Interpretation*, Pamphlets of the Stanford Literary Lab, Stanford 2017. Tillgänglig på https://litlab.stanford.edu/pamphlets/ (2017-10-13).
37 En sökning på ”elektrisk telegraf” i Språkbankens tidningskorpus visar dessutom att denna ordsammansättning blir mindre frekvent från och med 1850-talets slut. På 1860-talet skrivs det om ”telegraf” kort och gott, att den är elektrisk tas oftast för givet. Se https://spraakbanken.gu.se/korp/?mode=kubhist#?lang=sv&stats_reduce=word&cqp=%5B%5D&search=word%7Celektrisk%20telegraf&page=0&result_tab=4 (2017-1-13).
38 Vi har implementerat metoden i Chen & Lu (2011) som baseras på algoritmen i Lund & Burgess (1996). Se Zilong Chen & Yang Lu, ”A word co-occurrence matrix based method for relevance feedback”, *Journal of Computational Information Systems*, 2011:1, samt Kevin Lund & Curt Burgess, ”Producing high-dimensional semantic spaces from lexical co-occurrence”, *Behavior Research Methods, Instruments, & Computers* 1996:2. För att säkerställa att algoritmen implementerats på rätt sätt har den testats på samma textexempel som Chen & Lu utgått ifrån.
39 Vid testkörningar har fönsterstorleken vidgats och minimigränsen för samförekomsternas frekvenser sänkts. Testerna visade att de mest frekventa samförekomsterna fångas upp även när dessa parametrar förändras, men att svansen av lågfrekventa samförekomster – och brus – blir betydligt längre. För att minska bruset (svåridentifierade ordfragment) sattes minimigränsen vid sex förekomster och fönsterstorleken till fem ord, samma som Chen & Lu (2011) använder.
40 Patrice Flichy, *Dynamics of Modern Communication: The Shaping and Impact of New Communication Technologies*, London 1995, s. 37.
41 Flichy 1993, s. 34–39.
42 Se t.ex. Iwan Rhys Morus, ”’The nervous system of Britain’: Space, time and the electric telegraph in the Victorian age”, *The British Journal for the History of Science* 2000:4.
43 James W. Carey, ”Technology and ideology: The case of the telegraph”, i författarens *Communication as Culture: Essays on Media and Society*, Revised edition, New York & London 2008, s. 157.
44 Jonas Harvard, ”Connecting the Nordic region: The electric telegraph and the European news market”, i *Communicating the North: Media Structures and Images in the Making of*

the Nordic Region, Jonas Harvard & Peter Stadius (red.), Farnham & Burlington 2013, s. 48.

45 Se t.ex. Sarah Sharma, ”It changes time and space! Introducing power-chronography”, i *Communication Matters: Materialist Approaches to Media, Mobility and Networks*, Jeremy Packer & Stephen B. Croft Wiley (red.), London & New York 2012, s. 66.

46 Exemplen är hämtade från *Aftonbladet* 31/5 1845, 16/7 1846, 1/5 1852 och 24/2 1853, samtliga från tidningar.kb.se (2017-03-30).

47 ”Elektrisk telegraf mellan Europa och Förenta staterna”, *Aftonbladet* 21/7 1855, tidningar.kb.se (2017-03-30).

48 John Durham Peters, *Speaking into the Air: A History of the Idea of Communication*, Chicago & London 1999, s. 138f.

49 Simone M. Müller, ”From cabling the Atlantic to wiring the world: A review essay on the 150th anniversary of the Atlantic telegraph cable of 1866”, *Technology and Culture* 2016:3, s. 517f.

50 Harvard 2013, 47. Harvard har själv visat att detta antagande bör problematiseras, se ”Nya medier, gamla transporter. Hästar, tåg och ångbåt i den elektriska telegrafens tjänst”, i *1800-talets mediesystem*, Jonas Harvard & Patrik Lundell (red.), Stockholm 2010.

51 ”Elektriska telegrafen mellan England och Frankrike”, *Aftonbladet* 29/9 1851, tidningar.kb.se (2017-03-30).

52 David Thorburn & Henry Jenkins, ”Introduction: Toward an aesthetics of transition”, i *Rethinking Media Change: The Aesthetics of Transition*, Thorburn & Jenkins (red.), Cambridge & London 2003; Simone Natale & Gabriele Balbi, ”Media and the imaginary in history”, *Media History* 2014:2; Carolyn Marvin, *When Old Technologies Were New: Thinking About Electric Communication in the Late Nineteenth Century*, Oxford & New York 1988, s. 7f.

53 Rhys Morus 2000; Asa Briggs & Peter Burke, *A Social History of the Media: From Gutenberg to the Internet*, Cambridge & Malden 2002, s. 63f; James W. Carey & John J. Quirk , ”The mythos of the electronic revolution” i James W. Carey, *Communication as Culture: Essays on Media and Society*, Revised edition, New York & London, 2008.

54 Harvard 2010, s. 40.

55 Elihu Burritt, ”Tankar om krig”, *Aftonbladet* 13/5 1851, tidningar.kb.se (2017-03-30).

56 Exempelvis ”Föredrag af hr professorn Berg vid Svenska slöjdföreningens allmänna årssammankomstden 27 maj 1847”, publicerat i *Aftonbladet* 2/6 1847, en ledarartikel utan titel tryckt 6/6 1853, samt ett orubricerat brev från London 7/9 1854, samtliga från tidningar.kb.se (2017-03-30).

57 ”Diskussion öfver rektor Cramérs teser i läroverksfrågan”, *Aftonbladet* 7/12 1858, tidningar.kb.se (2017-03-30).

58 Johan Jarlbrink, ”Mobile/sedentary: News work behind and beyond the desk”, *Media History* 2015:3.

59 *Aftonbladet* 18/1 1855, tidningar.kb.se (2017-03-30).

60 Brian Winston, *Media, Technology and Society: A History from the Telegraph to the Internet*, London & New York 1998, s. 26–29.

61 Mussell 2013, s. 84.

62 Mussell 2012, s. 50–53.

63 Wolfgang Ernst, *Digital Memory and the Archive*, Minneapolis 2013, s. 99.