

Aftonbladet kunde i en notis i oktober 1847 rapportera att ny teknik inte alltid är att lita på när åskan fått den elektriska telegrafen att sända obehagliga textmeddelanden. Liknande brus har uppstått när just denna notis digitaliserats, men istället för åskan så är det nu den maskinella inläsningen (OCR) och den automatiska indelningen av textmassan till sammanhörande stycken som förvränger orden.

italisera.[1] Att arbeta med tidningstexter innebär i högre grad att finna metoder för att hantera det omfattande brus som tillförts materialet genom digitaliseringsprocessen.[2]

däribland den elektriska telegraphen. Syftet i denna poster är att beskriva arbetsprocessen, hur vi har gått till väga för att lokalisera och analysera de smutsiga texterna.

att han hvarken påminde sig de närmare omständigheterna vid slagsmålet, ej heller hade deruti kunnat deltaga.

Referenser

[1] Se t.ex. Matthew L. Jockers, *Macroanalysis: Digital Methods & Literary History* (Urbana 2013).

[2] Problemen beskrivs bl.a. i Traub, Myriam, Jacco von Osssenbruggen & Lynda Hardman, "Impact analysis of OCR quality on research tasks in digital archives", *Research and Advanced Technology for Digital Libraries*, 9316 (2015); Carolyn Strange, Daniel McNamara, Josh Wodak & Ian Wood, "Mining for the Future of Manuscript Research: The Impact of OCR Quality on the Use of Digitized Historical Newspapers", *Digital Humanities Quarterly* 2014:1.

[3] K. Lund and C. Burgess, Producing Highdimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203-208, 1996.

1. Sida i Aftonbladet 16/10 1847

[illegible]

Årsmäns,
 ingderna
 mänsk-
 eller
 brister
 och
 diade.

SJÖFÅRTS INNÄRTELSE.
Stockholms. Inkomne:
O. J. Ad. de Noire, Herrenska, bröder, Claes-
son, Carl, Hultén, Carl, Jönsson, J. O. Nord-
ström, Ad. Nilsson, Lindqvist, P. O. Nord-
ström, G. Örtengren, Carlsson Carl, Skjölde, O. O.
Carl, Löfdahl, Carl, Carlsson, Carl, Carlsson,
R. de Hassel, Persson, Nordmann, O. de Haan,
Lundström, R. Larsson, Sörensen, H. Holm, Ab-
dullah, Carl, Carlsson, Carl, Carlsson, Carl,
de Kapt. Torvald, H. Carlsson.

[illegible]

ER. I dag Lördag ugifres 36 42 af
Läsebibliotheket,
Feskoakrifst och Magasin för den sköna Lit-

Open	Power	Thermostat	Control	Set	Unit	Max	Min	OK	Cancel
TELEGRAM	23.00°C	23.00°C	23.00°C	23.00°C	23.00°C	23.00°C	23.00°C	23.00°C	23.00°C
On	Off	On	Off	On	Off	On	Off	On	Off

Målet är att beräkna samförekomst av ord för alla stycken i tidningstexten där konceptet ELEKTRISK TELEGRAF förekommer. Vi måste därför finna alla "stavningsvarianter" av ELEKTRISK och TELEGRAF med hänsyn taget till bruset. Alla unika ord i materialet med ett Levenshtein distance av 2 eller mindre från konceptens grundformer identifieras. Resultatet är en bruttolista med ord som därefter rensas manuellt.

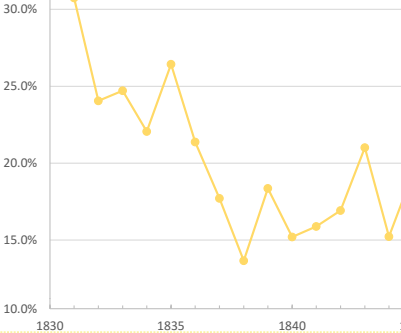
I det undersökta materialet återfanns 4727 ord med avståndet 2 eller mindre till TELEGRAF av vilka 3524 ord kvarstod efter den manuella genomgången.

Mängden stavningsvarianter används för att identifiera relevanta stycken i material där båda koncepten förekommer (ELEKTRISK och TELEGRAF). Vi utgår från de textblock som digitaliseringsprocessen skapat. Denna indelning innehåller dock stora brister genom att stycken både slagits samman och splittrats på ett felaktigt sätt. Det sker därför en manuell genomgång av alla textblock där rader som felaktigt sammanslagits med stycken som handlar om aktuella koncept sorteras bort.

**Totalt återfanns 1253 textblock om totalt 124205 rader text.
Efter manuell genomgång återstod 22162 rader text.**

För att hitta mönster i hur ELEKTRISK TELEGRAF framställs i materialet beräknas samförkomst enligt Hyperspace Analogue to Language (HAL) [3]. Analysen sker per decennium för att spåra förändringar över tid. Beräkningen baseras på hur ofta två ord förekommer tillsammans i ett fönster av 5 ord inom samma stycke. Fönstret flyttas stegvis genom hela stycket och beräkningen upprepas för varje ord i stycket. För varje ord i varje ord i h hela materialet. En list med stopword används för att filtrera bort mindre betydelsebärande ord. För att begränsa antalet ordpar i resultatet har vi filtrerat bort de som förekommer i färre än sex fönster.

är bekant och hafta omtalats att äskan stundom inblandar sig i saker på detta slag meddela franska tidningarna följande bref till elektriska telegraferna i Frankrike »**I ondags** efter kl 5 under ett jernvägen till S:t Germain Stationens telegrafist trodde att en rapp men som deras sammanhang icke kunde utelas ärnade han gifva ö raten fick ett starkt elektriskt slag följdyt af en knall likasom af ett metallträdarne som tjena elektriciteten till de **visl** verdärgidav **visl** erslagna ken —tas till 70 70 misvärt fruktarsnart **ta** ef och sisrans n



A horizontal timeline with a central axis. The year 1835 is marked on the left, and 1840 is marked further to the right. There are tick marks between these years, indicating annual intervals.

Fördelen med metoden är att även förvårdade ordvarianter tas med i resultatet. Att bestämma minimigränsen för hur många gånger ett ord eller så förekomma utgör dock ett dilemma. Med ett högre värde filtrerar man bort ord som inte förekommit tillräckligt ofta för att vara av betydelse för felaktig autosegmentering. Men då riskerar man även att missa ovanliga förekomningar av vanliga ord. Med ett femdofersdörrslag och en minimifrekvens på sex förekomster hittades 628 ord som sannolikt i 1860 ordpar som inte hade varit tillräckligt vanliga för att kunna användas i ordpar. Ord som förvärrats av OCR-tekniken och som förekommer tillräckligt ofta för att komma med är bland annat "oceanen", "olecriska", "Stoeholm" och "ëg:Stoeholms". En manuell genomgång gör så att dessa kan slås samman med de korrekta motsvarigheterna. Därefter räknas ut de korrekta svarade resultatet i Gephi.