



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E

## Article Separation

### Contact:

Roger Labahn

✉ roger.labahn@uni-rostock.de

### LÉON XIII

*«On ne pouvait pas s'attendre à ce que le pape, qui n'est pas un homme d'État, ait une telle responsabilité en Italie.»*

Rome, 15 mars.

Monseigneur le cardinal en chef.

Le *Matin* publie, à 9 h, quelques jours, un télégramme de son correspondant de Rome, éclairant certains recules de l'histoire récente des rapports entre le Vatican et le Quirinal.

«Permettez-moi de venir à mon tour mettre sous les yeux de nos lecteurs tout un chapitre de cette histoire, en vous garantissant son exactitude même dans les plus minimes détails.»

«Avant de prendre place au Sacré Collège, Léon XIII était évêque de Perugia, où il avait sa séquelle toute les sympathies. Tandis que Mgr. Poci, commandant en chef de Perugia, un général de grand talent et de beaucoup d'expérience, le général Cardini commandait la division militaire, dont le siège était tenu à côté de l'évêché. Les deux chefs eurent occasion de se rencontrer et de s'apprécier mutuellement — ils échangeaient des visites — bref, ils devinrent bons amis.»

«L'évêque avait été nommé cardinal et le général Cardini en 1871, avant sa loi d'abolition d'un Pape. Le conflit entre le gouvernement italien et le Vatican devenait de jour en jour plus aigu, et les deux amis, absolument italiens, en déplorèrent les conséquences.»

«Le général, dont l'élégance s'alliait à une grande culture, ne fut pas oublié par les vives positions du boulevard, où il se tenait pendant quelques années, discutant tout le monde par son

sensuel tiré du château. Saint-Angel, pour saluer le nom de Léon, qui le gouvernement italien avait officiellement reconnu.»

«Le Pape, après avoir reconnu à son tour le royaume d'Italie, avec Rome capitale, et devant sa félicitation aux troupes du roi.»

«Il y avait encore une chose, qui n'intéressait ni la cour, ni le peuple, mais qui intéressait beaucoup le cardinal di Pietro — elle n'était pas bien longue — la voici.»

«Le roi payait les dettes du cardinal di Pietro.»

«Le général Cardini avait obtenu l'adhésion du cardinal Poci au plan arrangé entre Victor Emmanuel et di Pietro, mais, malheureusement, tout se termina par un échec.»

«On commença par payer les dettes du cardinal — ce qui n'était pas une petite affaire — mais le roi donna l'ordre, en suite du Parlement obligé M. Nicotri à donner sa démission de ministre de l'Intérieur.»

«Le personnel de police spéciale, qui avait conduit les négociations avec le Vatican, se retira avec le ministre, ne voulant pas servir son successeur, M. Crispien, dans lequel on avait une grande confiance.»

«Le roi, qui n'était pas non plus M. Crispien, se gela bien de le mettre en contact de sa politique papale.»

«Il attendit le retour de M. Nicotri au pouvoir. En 1878, avant sa mort de 80 ans, on n'était pas très pressé.»

«Mais les événements se précipitèrent — Victor Emmanuel tomba malade et fut emporté presque subitement, sans pouvoir donner ses instructions à son fils — et sans informer ses ministres de ce qu'il avait préparé.»

«Quelques semaines après, le Pape mourut

### APRÈS BAC-NINH

#### LES CONSÉQUENCES DE LA VICTOIRE

#### A LA POURSUITE DE L'EN-NEMI

**Dispersion des Vietnams Français — Les Chinois au Nord — Les Français au Sud — La France.**

(par son auteur)

Le 15 mars, 15 mars — On mande de Hanoi, 14 mars, 7 h 15 du matin —

«La colonne du général Négrier est entrée à Bac-Ninh dans la soirée du 12.»

«L'ennemi dans la retraite de se voir couper la retraite, s'est enfui vers la frontière, laissant la colonne de derrière.»

«Les Français, pendant leur marche en avant, n'ont éprouvé aucune faible résistance.»

«Leurs pertes sont insignifiantes.»

«Leur succès est complet.»

Marche en Avant.

(par son auteur)

Le 15 mars, 15 mars — On télégraphie de Hanoi, 14 mars, 14 mars — La moitié de la brigade du général Millot a attaqué et enlevé cinq forts et quatre villages sur la route stratégique de Bac-Ninh.

«Les Chinois possèdent des fortifications admirablement construites et armées sur les collines, mais ils n'ont opposé aucune résistance.»

tenue en Tonkin a, un jour, montré de toute sa puissance. Déjà, cette affaire n'est qu'un demi-succès, selon le mot du *Nation*. Tel est, ainsi, l'avis de nos amis, constaté par le *Journal*. *Paris*, et par le *Lézard*, qui conseille sa gouvernement d'attendre les avis de la population et de la région.»

Marche des Colonnes Millot et Négrier.

(par son auteur)

Le 15 mars, 15 mars — Le correspondant du *Times*, attaché à la colonne du général Millot, a télégraphié, à la date du 13. Après une marche difficile de quatre jours à travers des champs de riz, étaient les routes directes, nord et sud du canal des Rapides, routes qui étaient précédemment fortifiées, les troupes françaises ont passé le canal à un endroit au sud-est de Bac-Ninh où précédemment il y avait déjà un pont.

«Le général Négrier a pris aujourd'hui une forte position sur les hauteurs. Une faible résistance lui a été opposée — les Français n'ont subi aucune perte — La santé et l'esprit des troupes sont excellents. Les Chinois sont au nombre de 2.000 environ. La poursuite se continue.»

Les Morts et Blessés de «Xacouang».

«Voici les noms des hommes de la compagnie de Montigny, tués ou blessés, dans la journée du 12 mars, au Tonkin, sous les ordres de leur commandant.»

Diazul, médecin, tué.  
Léall, second maître, blessé.  
Bénistat, maître, blessé.

Article Separation (AS) is defined as the task of identifying regions on a newspaper page that hold information to one article and is an important step for further natural language processing applications in the information retrieval pipeline.

However, the AS task can be divided into two parts: (text) block recognition and subsequent clustering.

Since we already determine baselines/textlines as a preprocessing step for the Automated Text Recognition, one way to describe these text blocks is by clusters of baselines as shown in the figure.

To achieve this, our research focuses mainly on algorithms based on machine learning in addition to the use of traditional clustering methods.



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

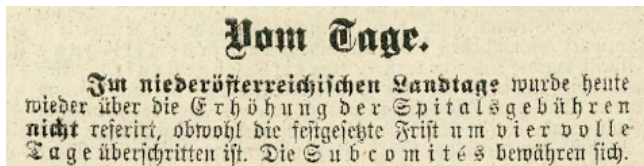


# Automated Text Recognition

## Contact:

Roger Labahn

✉ [roger.labahn@uni-rostock.de](mailto:roger.labahn@uni-rostock.de)



## Vom Tage.

Im niederösterreichischen Landtage wurde heute wieder über die Erhöhung der Spitalsgebühren nicht referirt, obwohl die festgesetzte Frist um vier volle Tage überschritten ist. Die Subcomités bewähren sich.

*Automatic Text Recognition provides a transcription for each text line*

Automatic Text Recognition (ATR) is the task of automatically transcribing the textual contents contained in images. This is a fundamental step in the NewsEye workflow, as it provides the textual data which is needed to further be processed in Natural Language Processing tasks.

In our case, we apply models on segmented text line images, which are the results of a Layout Analysis algorithm. These models involve deep Artificial Neural Networks designed for sequence labeling tasks. Additionally, the results of such ATR models include posterior character probabilities, which can be used to search in the transcription by means of keyword spotting.



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu



# NewsEye Demonstrator

## Contact:

Antoine Doucet  
antoine.doucet@univ-lr.fr



The NewsEye demonstrator (<https://platform.newseye.eu>) showcases the different outputs of the project. Its website, freely available, allows users to explore newspaper collections in different languages. Ranging from 1850 to 1950, these newspapers were made available by three National Libraries

of the project in Vienna, Paris and Helsinki. After searching through the collection, users can create personal collections to gather documents of interest. These collections can be exported in various formats or analysed directly inside the Demonstrator.



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu



# Semantic Text Enrichment for OCR'd content

## Contact:

Antoine Doucet

✉ antoine.doucet@univ-lr.fr

Semantic text enrichment consists of analyzing documents and adding semantic metadata to their contents. Among several semantic metadata, we concentrate on named entities and, more precisely, we aim to recognize these entities in documents and disambiguate them to a knowledge base. In addition, these entities will be associated with the stance of the text in which they are mentioned.

However, OCR'd documents contain numerous errors due to the state of documents, as a result of aging, poor storage conditions and/or the low quality of initial printing materials. These errors reduce the performance of all downstream natural language processing tasks (such as NER, NEL and stance detection).

To overcome these problems, NewsEye designs approaches based on the most recent techniques (neural networks and deep learning) that are robust to OCR problems and language independent. Out of 13

competitors, NewsEye approaches reached first place in 50 of 52 leaderboards on NER and NEL in English, French and German at the CLEF HIPE 2020 competition.

## Reasonable quality

Her name is **Clare** she is 25. She has one cat and her name is **George**. His has one sister. She lives in **England**. She has a boy friend and his name is **Joe**. Her birthday is on **17 September 1988**. Her favourite TV series is **Grey's anatomy**. Her favourite movie is **Cartoons**. Her favourite school subject is **English** and her favourite sport is **swimming**. Her favourite hobby is **eating** and her favourite country is **England**.

## Medium quality

Her name is **Clare**, he is 25. She has one cat and her name is **George**. His has one sister. She lies in **England**. She has a boy friend and his name is **Joe**. Her birthday is on **17 September 1**. Her favourite TV series is **Greys anatomy**. Her favourite movie is **Cartoons**. Her favourite sport is **swimming**. Her favourite hobby is **eating** and her favourite country is **England**.

## Poor quality

Her name is **Clare**, she is 25. She has one cat and her name is **George**. His has one sister. She lies in **England**. She has a boy friend and his name is **Joe**. Her birthday is on **17 September 1**. Her favourite TV series is **Greys anatomy**. Her favourite movie is **Cartoons**. Her favourite sport is **swimming**. Her favourite hobby is **eating** and her favourite country is **England**.



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)  [@newseyeeu](https://twitter.com/newseyeeu)

vote, the right to wear pants), as well as the female figures who infiltrate the journalistic space: female writers/journalists. What place is given to them in the newspapers? What role do newspapers play in the struggles of women? Digital tools (notably topic modeling and frequency tools) are very valuable in this case study for revealing the trends and key figures.



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Case Study: Media & Journalism

The interaction  
between editorial  
office and readers

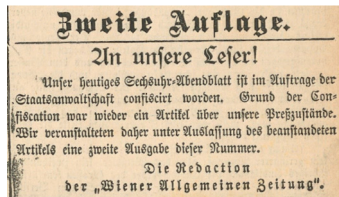
### Contact:

Eva Pfanzelter

✉ Eva.Pfanzelter@uibk.ac.at

## An die künftigen Abonnenten.

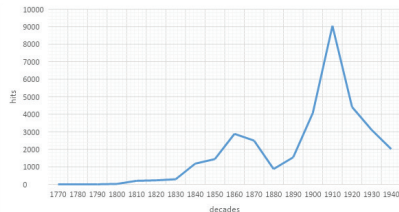
Hast du nie ein Schiff vom Stapel laufen sehen, lieber Leser? Es ist ein interessanter Anblick, und gar manche Betrachtungen und Gedanken, heiz



Wiener Allgemeine Zeitung, 13 November 1880, page 1, evening paper

In the “long” 19th century, newspapers became the mass medium par excellence. At the same time, this created a new public sphere, where the negotiations of the modern nation state played an important role. However, the common perception that the dissemination

of news and opinion by the press during this period was a one-way, linear process in which information was carried from the editorial office to the readers is not entirely correct. There were different possibilities of interaction and two-way communication. This case study therefore focuses on this interaction.



Absolute frequency of the phrase „an die Redaktion“ („to the editors“) per decade



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Case Study on Nationalism

The language of Nationalism  
1850-1950

### Contact:

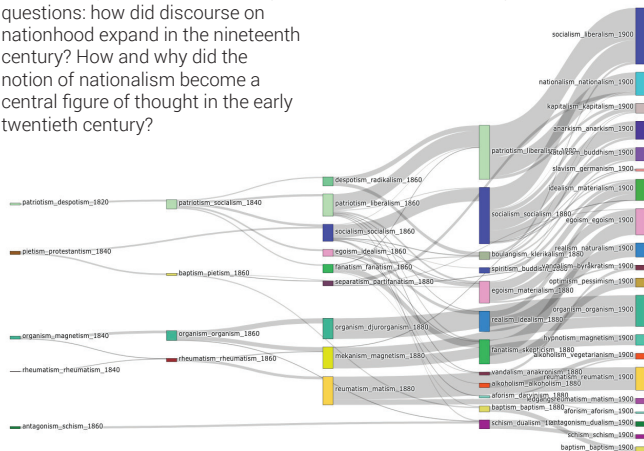
Jani Marjanen

✉ jani.marjanen@helsinki.fi

*Finska Språket, betraktadt såsom Nationalspråk.*

The case study on nationalism focuses on the changing language of nationhood in the late 19th and early 20th century. It asks the following questions: how did discourse on nationhood expand in the nineteenth century? How and why did the notion of nationalism become a central figure of thought in the early twentieth century?

The case combines qualitative reading of sources with frequency analysis and word embedding models to cluster key terms.



Clusters 1820-1900



## A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 [@newseyeeu](https://twitter.com/newseyeeu)

## Case Study on Migration Return Migration 1850-1950

### Contact:

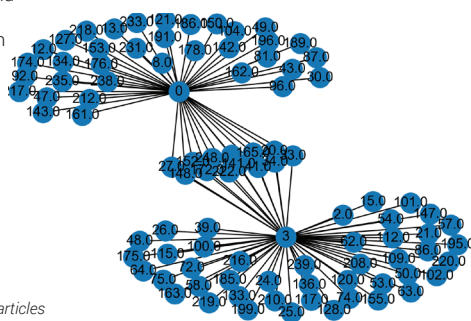
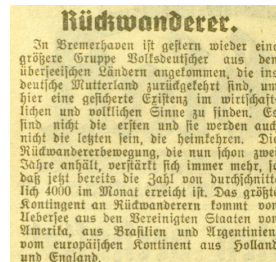
Sarah Oberbichler

✉ [sarah.oberbichler@uibk.ac.at](mailto:sarah.oberbichler@uibk.ac.at)

**Sie sind da, sie sind da!**

The case study on migration attempts to reveal return migration processes to Europe between 1850 and 1950. It does so by asking the following questions: how and in what context were Austrian daily newspapers reporting on returnees and how did the reporting change over time?

The case uses digital methods like topic modeling and word embedding to create a subcollection for further qualitative analysis (discourse and argumentation analysis).



Topic visualization for  
relevant and non-relevant articles



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

✉ [roger.labahn@uni-rostock.de](mailto:roger.labahn@uni-rostock.de)

~~Déraud, mécanicien, tué.~~  
~~Lesall, second-maitre, blessé.~~  
~~Réaillant, matelot, blessé.~~

tunnistaminen ja sitä seuraava ryvästys. Tekstialueet voidaan esittää automaattisen tekstintunnistuksen tuottamien tekstirivien ryppäinä, kuten kuvassa, ja artikkelit edelleen niitten ryppäinä. Tätä varten tutkimuksemme keskittyy perinteisten klusterointialgoritmien lisäksi pääosin koneoppimiseen perustuviin algoritmeihin.



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

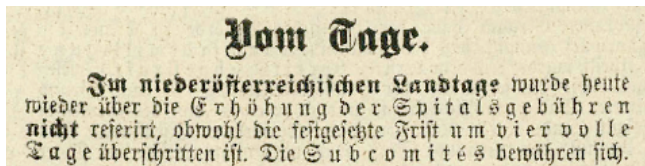


# Automaattinen tekstin- tunnistus

## Lisätietoja:

Roger Labahn

✉ roger.labahn@uni-rostock.de



## Vom Tage.

Im niederösterreichischen Landtage wurde heute wieder über die Erhöhung der Spitalsgebühren nicht referirt, obwohl die festgesetzte Frist um vier volle Tage überschritten ist. Die Subcomités bewähren sich.

*Automaattinen tekstintunnistus tekee transkription jokaisesta tekstirivistä*

Automaattinen tekstintunnistus (Automatic Text Recognition, ATR) muuttaa kuvien sisältämän kirjoituksen tekstimuotoon. Se on keskeinen vaihe NewsEyen työnkulussa, koska siinä muodostetaan tekstimuotoinen data, jota tarvitaan luonnollisen kielen käsittelyprosesseissa.

Käyttämässämme menetelmässä

malleja sovelletaan riveittäin jaettuihin kuviin, jotka rakenneanalyysialgoritmi on muodostanut. Mallit hyödyntävät sekvenssien tunnistamiseen suunniteltuja syväoppivia neuroverkkoja. Lisäksi ATR-mallien tulokset sisältävät merkkikohtaiset tunnituksen todennäköisyydet, joita voidaan hyödyntää avainsanojen tunnistukseen perustuvissa hauissa.



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu



# NewsEye Demonstrator

## Lisätietoja:

Antoine Doucet

✉ antoine.doucet@univ-lr.fr



NewsEye Demonstrator (<https://platform.newseye.eu>) esittelee projektin tuotoksia. Verkossa vapaasti käytettävissä olevalla sivustolla käyttäjät voivat tarkastella erikielisiä sanomalehtikokoelmia, jotka ulottuvat ajallisesti vuodesta 1850 vuoteen 1950. Sanomalehdet on saatu käyttöön projektiin

osallistuvilta Itävallan, Ranskan ja Suomen kansalliskirjastoilta. Tekemiensä hakujen pohjalta käyttäjät voivat muodostaa itselleen kokoelmia heitä kiinnostavista sisällöistä. Näitä kokoelmia on mahdollista tallentaa eri tiedostomuodoissa tai analysoida suoraan Demonstratorissa.



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu



# Semanttinen tekstin rikastaminen

Automaattisesti  
tunnistetun tekstin

## Lisätietoja:

Antoine Doucet

✉ antoine.doucet@univ-lr.fr

Semanttinen tekstin rikastaminen tarkoittaa asiakirjojen analysointia ja semanttisten metatietojen lisäämistä niiden sisältöön. Lukuisten semanttisten metatietojen joukosta keskityimme nimettyihin kohteisiin. Tarkemmin sanottuna pyrimme tunnistamaan nimetyt kohteet asiakirjoissa ja liittämään ne yksiselitteisesti tietämyskantaan. Kohteisiin liitetään myös tieto, miten niihin tekstissä asennoidutaan.

Koneellisesti tunnistetussa tekstissä voi olla runsaasti virheitä. Tämä johtuu ajan ja huonojen säilytysolosuhteiden vaikutuksesta asiakirjojen kuntoon ja/tai alkuperäisten painomateriaalien heikosta laadusta. Virheet heikentävät kaikenlaisten luonnollisen kielen rikastamisen tapojen tuottamaa laatua.

NewsEye suunnittelee uusimpiin tekniikoihin (neuroverkkoihin ja syväoppimiseen) perustuvia menetelmiä, jotka eivät ole herkkiä tekstintunnistusvirheille ja jotka ovat kielestä riippumattomia.

## Kohtuullinen laatu

## Keskittason laatu

## Huono laatu



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

käyttää housuja) sekä naisista, jotka tunkeutuvat journalismin alueelle: naispuolisista kirjoittajista/ journalisteista. Millainen paikka heille annetaan sanomalehdissä? Millainen rooli sanomalehdillä on naisten kamppailuissa? Digitaaliset työkalut (etenkin aihemallinnus ja frekvenssi työkalut) ovat erittäin hyödyllisiä trendien ja avainhenkilöiden paljastamisessa.



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Median & journalismin tapaustutkimus

Toimituksen ja  
lukijoiden  
vuorovaikutus

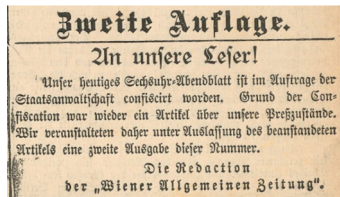
### Lisätietoja:

Eva Pfanzelter

✉ Eva.Pfanzelter@uibk.ac.at

### An die künftigen Abonnenten.

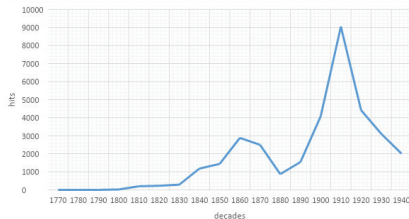
Haft du nie ein Schiff vom Stapel laufen sehen, Lieber Leser? Es ist ein interessanter Anblick, und gar manche Betrachtungen und Gedanken, heiz



Wiener Allgemeiner Zeitung, 13. marraskuuta, 1880, sivu 1, iltapäivä

Sanomalehdet muodostuivat ns. pitkän 1800-luvun aikana massamediaiksi jonkalaista ei ennen ollut olemassa. Tämä muokkasi julkisuutta, jossa määrittely kansakunnasta oli tärkeässä roolissa. Melko yleinen käsitys siitä, että uutisia välitettiin

yksisuuntaisesti toimitukselta lukijoille, ei pidä täysin paikkaansa. Lehdissä oli käytössä monia keinoja kaksisuuntaiselle kommunikaatiolle. Tämä tapaustutkimus keskittyy tähän vuorovaikutukseen.



"Toimitukselle"-fraasin ("an die Redaktion") absoluuttinen frekvenssi vuosikymmentä kohden



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Nationalismin tapaustutkimus

Nationalismin kieli  
1850-1950

### Lisätietoja:

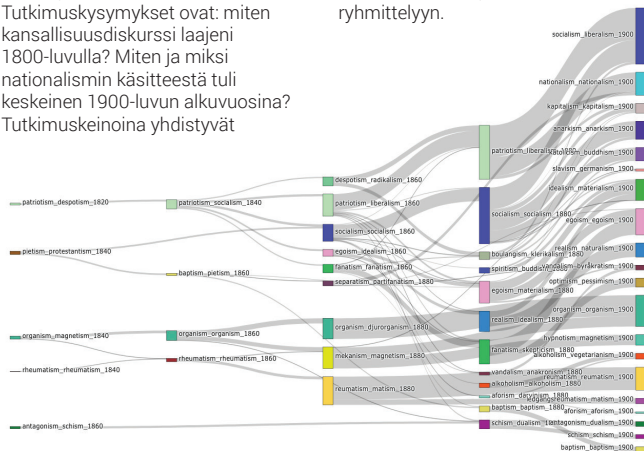
Jani Marjanen

✉ jani.marjanen@helsinki.fi

*Finska Språket, betraktadt såsom Nationalspråk.*

Tämä tapaustutkimus keskittyy kansakuntaa koskevan kielen muutokseen 1800-luvun lopulla ja 1900-luvun alussa. Tutkimuskysymykset ovat: miten kansallisuuskysymys laajeni 1800-luvulla? Miten ja miksi nationalismin käsitteestä tuli keskeinen 1900-luvun alkuvuosina? Tutkimuskeinoina yhdistyvät

lähteiden laadullinen lukeminen ja kvantitatiivinen frekvenssianalyysi. Lisäksi tutkimuksessa käytetään sanaopetusmalleja avaintermien ryhmittelyyn.



Klusterit 1820 – 1900



## A Digital Investigator for Historical Newspapers

Tässä projektissa humanististen tieteiden tutkijat, tietojenkäsittelytieteen tutkijat ja kirjastojen asiantuntijat työskentelevät yhdessä kansalliskirjastojen (Itävalta, Ranska, Suomi) historiallisten sanomalehtien parissa.

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Siirtolaisuuden tapaustutkimus

Paluumuutto  
1850-1950

### Lisätietoja:

Sarah Oberbichler

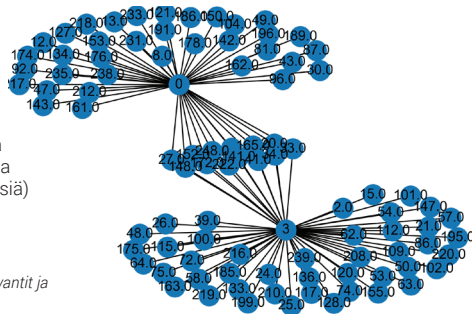
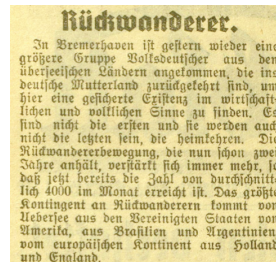
✉ [sarah.oberbichler@uibk.ac.at](mailto:sarah.oberbichler@uibk.ac.at)

**Sie sind da, sie sind da!**

Siirtolaisuuden tapaustutkimus pyrkii paljastamaan Eurooppaan suuntautuneen paluumuuton prosesseja vuosina 1850 – 1950. Tutkimuksessa kysytään: kuinka ja missä asiayhteydessä itävaltalaiset päivälehdet tiedottivat paluumuuttajista, ja miten tiedottaminen muuttui ajan kuluessa?

Tutkimuksessa käytetään digitaalisia menetelmiä kuten aihemallinnusta ja sanaopusmalleja. Näiden avulla luodaan osakokoelma tarkempaa laadullista analyysiä (diskurssi- ja argumentaatioanalyysiä) varten.

*Aiheiden visualisointi: relevantit ja epärelevantit artikkelit*





## A Digital Investigator for Historical Newspapers



Dans le cadre de ce projet, des chercheurs en sciences humaines, des informaticiens et des bibliothécaires travaillent ensemble sur la presse historique conservée dans trois bibliothèques nationales (Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

## Segmentation d'article

✉ [roger.labahn@uni-rostock.de](mailto:roger.labahn@uni-rostock.de)

Départ, mécanicien, fus.  
Lesall second-maitre, blessé.  
Résistant matelot blessé.

nous avons déjà déterminé les lignes de texte comme une étape préparatoire du traitement pour la reconnaissance automatique de texte, l'on peut décrire ces blocs de texte comme des regroupements de lignes de texte tel que l'illustre cette image. Pour atteindre ce résultat, notre recherche repose principalement sur des algorithmes basés sur l'apprentissage machine, en plus de méthodes de regroupement classiques.



## A Digital Investigator for Historical Newspapers



Dans le cadre de ce projet, des chercheurs en sciences humaines, des informaticiens et des bibliothécaires travaillent ensemble sur la presse historique conservée dans trois bibliothèques nationales (Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

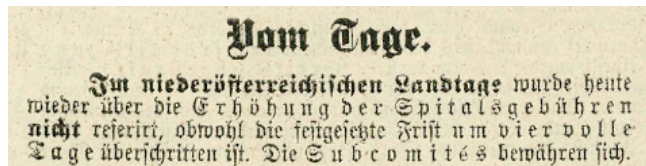


## Reconnaissance automatique du texte

### Contact :

Roger Labahn

✉ [roger.labahn@uni-rostock.de](mailto:roger.labahn@uni-rostock.de)



### Vom Tage.

Im niederösterreichischen Landtage wurde heute wieder über die Erhöhung der Spitalsgebühren nicht referirt, obwohl die festgesetzte Frist um vier volle Tage überschritten ist. Die Subcomités bewähren sich.

*La reconnaissance automatique du texte fournit une transcription pour chaque ligne de texte.*

La reconnaissance automatique du texte consiste à transcrire automatiquement les contenus textuels d'une image. C'est une étape fondamentale dans le flux de travail de NewsEye car elle fournit les données textuelles nécessaires pour être ensuite traitées dans des opérations de traitement automatique du langage naturel. Dans notre cas de figure, nous appliquons des modèles de reconnaissance à des images de

lignes de textes segmentées qui sont le produit d'un algorithme d'analyse de la mise en page. Ces modèles impliquent des réseaux de neurones artificiels profonds conçus pour des tâches d'étiquetage de séquences. De plus, les résultats de ces modèles incluent des probabilités appliquées à la reconnaissance de chaque caractère, qui peuvent être utilisées pour chercher dans les transcriptions au moyen de techniques de repérage de mots-clés (*keyword spotting*).



## A Digital Investigator for Historical Newspapers



Dans le cadre de ce projet, des chercheurs en sciences humaines, des informaticiens et des bibliothécaires travaillent ensemble sur la presse historique conservée dans trois bibliothèques nationales (Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

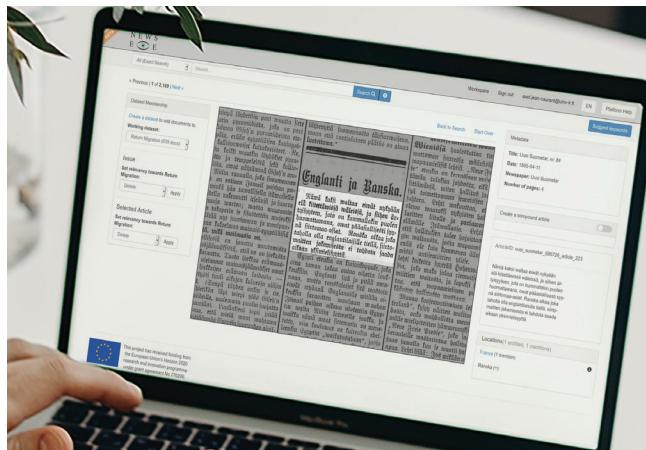


# Le démonstrateur NewsEye

## Contact :

Antoine Doucet

✉ antoine.doucet@univ-lr.fr



Le démonstrateur NewsEye (<https://platform.newseye.eu>) présente les résultats du projet. Ce site, dont l'accès est gratuit, permet à l'utilisateur d'explorer des collections de presse dans différentes langues. Publiés entre 1850 et 1950, ces journaux ont été mis à disposition

par les trois bibliothèques nationales basées à Vienne, Paris et Helsinki. A l'issue des recherches, l'on peut constituer des corpus personnalisés correspondant à des centres d'intérêt. Ceux-ci peuvent être exportés sous divers formats ou analysés au sein même du démonstrateur.



A Digital Investigator for  
Historical Newspapers



Dans le cadre de ce projet, des chercheurs  
en sciences humaines, des informaticiens  
et des bibliothécaires travaillent ensemble  
sur la presse historique conservée dans  
trois bibliothèques nationales  
(Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu



## Enrichissement sémantique de contenus océrés

### Contact :

Antoine Doucet

✉ antoine.doucet@univ-lr.fr

L'enrichissement sémantique de corpus textuels consiste en l'analyse de documents et l'ajout de données sémantiques à leurs contenus. Parmi une palette de données sémantiques, nous nous concentrons sur les entités nommées et, plus précisément, nous visons à reconnaître ces entités dans les documents, à les désambiguïser via une base de connaissances. De plus, les extraits de texte dans lesquels elles figurent leur sont associés.

Cependant, le texte des documents océrés contient de nombreuses erreurs, du fait du vieillissement des documents, de leurs conditions de stockage et/ou de la qualité médiocre du support d'impression originel. Ces erreurs réduisent les performances de tous les traitements ultérieurs en langage naturel (tels que la reconnaissance des noms de personnes, de lieux, d'organisations, et l'analyse d'opinion et de sentiment).

Afin de surmonter ces difficultés, NewsEye conçoit des approches basées sur les techniques les plus récentes (réseaux de neurones artificiels et apprentissage profond) résistantes aux erreurs de la reconnaissance automatique du texte et indépendantes de tout langage. Sur 13 concurrents,

les méthodes développées pendant le projet NewsEye ont atteint la première place dans 50 des 52 classements pour la reconnaissance des noms de personnes, de lieux et d'organisations en anglais, français et allemand au concours CLEF HIPE 2020.

### Qualité raisonnable

### Qualité moyenne

### Mauvaise qualité



A Digital Investigator for  
Historical Newspapers



Dans le cadre de ce projet, des chercheurs  
en sciences humaines, des informaticiens  
et des bibliothécaires travaillent ensemble  
sur la presse historique conservée dans  
trois bibliothèques nationales  
(Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

pantalon) ainsi qu'aux personnalités féminines qui investissent la sphère journalistique. Quelle est la place faite aux femmes dans les journaux ? Quel est le rôle de la presse dans les luttes des femmes ? Les outils numériques (en particulier la modélisation thématique et les outils d'analyse de fréquence) sont particulièrement précieux dans ce cas d'étude pour révéler les tendances et les éléments clés.



A Digital Investigator for  
Historical Newspapers



Dans le cadre de ce projet, des chercheurs en sciences humaines, des informaticiens et des bibliothécaires travaillent ensemble sur la presse historique conservée dans trois bibliothèques nationales (Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

## Étude de cas : médias et journalisme

L'interaction entre  
la rédaction et les  
lecteurs

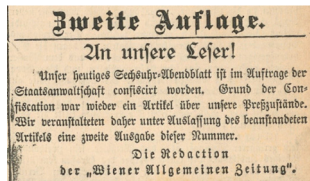
### Contact :

Eva Pfanzelter

✉ Eva.Pfanzelter@uibk.ac.at

## An die künftigen Abonnenten.

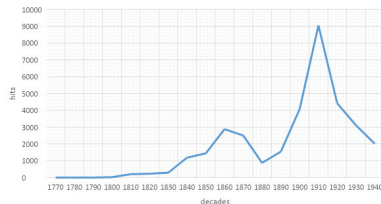
Hast du nie ein Schiff vom Stapel laufen sehen, lieber Leser? Es ist ein interessanter Anblick, und gar manche Betrachtungen und Gedanken, heiz



Wiener Allgemeine Zeitung, 13 novembre 1880, page 1, édition du soir

Pendant le « long » 19ème siècle, la presse est devenue le média de masse par excellence. Dans le même temps, un nouvel espace public dans lequel les négociations autour de l'État-nation moderne ont joué un rôle important, a vu le jour. Cependant, la croyance populaire selon

laquelle la diffusion d'informations et d'opinions dans la presse est un processus linéaire et à sens unique, où l'information n'est transmise que des éditeurs aux lecteurs, n'est pas entièrement vraie. Différents types d'interactions et de communications existaient. Cette étude de cas se concentre par conséquent sur ces interactions.



Fréquence absolue de l'expression "an die Redaktion" ("pour les éditeurs") par décennie



## A Digital Investigator for Historical Newspapers



Dans le cadre de ce projet, des chercheurs en sciences humaines, des informaticiens et des bibliothécaires travaillent ensemble sur la presse historique conservée dans trois bibliothèques nationales (Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Etude de cas sur le nationalisme

Le langage du nationalisme 1850-1950

### Contact :

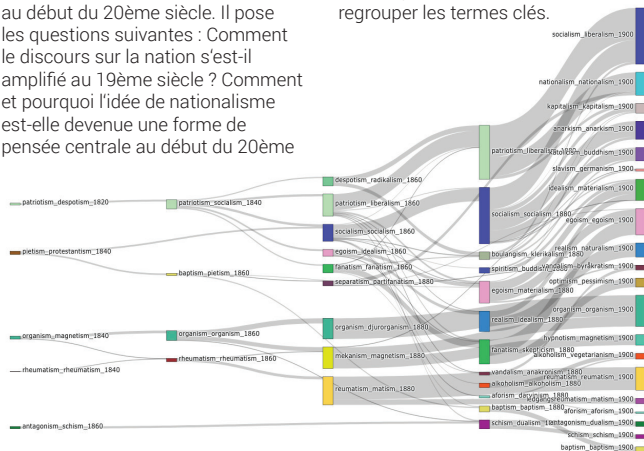
Jani Marjanen

 jani.marjanen@helsinki.fi

*Finska Språket, betraktadt såsom Nationalspråk.*

Ce cas d'étude consacré au nationalisme met l'accent sur le changement linguistique du concept de nation à la fin du 19ème siècle et au début du 20ème siècle. Il pose les questions suivantes : Comment le discours sur la nation s'est-il amplifié au 19ème siècle ? Comment et pourquoi l'idée de nationalisme est-elle devenue une forme de pensée centrale au début du 20ème

siècle ? Une lecture qualitative des sources est combinée avec une analyse de fréquence et des modèles de plongement lexical destinés à regrouper les termes clés.



Clusters 1820-1900



A Digital Investigator for  
Historical Newspapers



Dans le cadre de ce projet, des chercheurs  
en sciences humaines, des informaticiens  
et des bibliothécaires travaillent ensemble  
sur la presse historique conservée dans  
trois bibliothèques nationales  
(Autriche, Finlande, France).

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

## Etude de cas sur les migrations

Migrations de retour  
1850-1950

Contact :

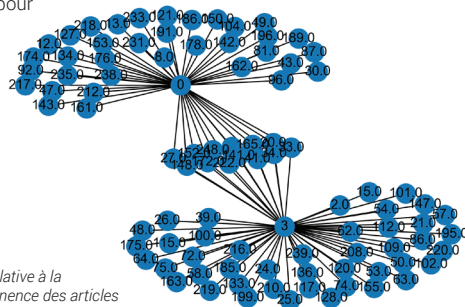
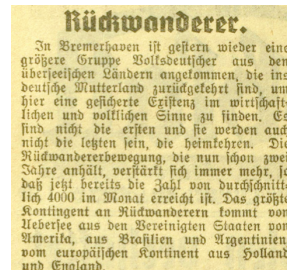
Sarah Oberbichler

✉ [sarah.oberbichler@uibk.ac.at](mailto:sarah.oberbichler@uibk.ac.at)

## Sie sind da, sie sind da!

Ce cas d'étude sur les migrations tente de mettre en lumière ce phénomène migratoire vers l'Europe entre 1850 et 1950. Pour ce faire, les questions suivantes sont posées : Comment et dans quel contexte les quotidiens autrichiens ont-ils évoqué les rapatriés, et comment le discours a-t-il évolué au fil du temps ?

Le cas d'étude utilise des méthodes numériques comme la modélisation thématique, le plongement lexical pour constituer une sous-collection pour des analyses qualitatives ultérieures (analyse du discours et des arguments).



Visualisation thématique relative à la  
pertinence ou à la non pertinence des articles



A Digital Investigator for  
Historical Newspapers

In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

✉ [roger.labahn@uni-rostock.de](mailto:roger.labahn@uni-rostock.de)



A Digital Investigator for  
Historical Newspapers

In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

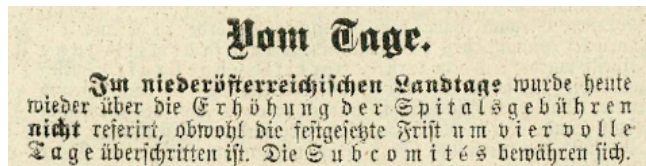


# Automatische Text- erkennung

## Kontakt:

Roger Labahn

✉ [roger.labahn@uni-rostock.de](mailto:roger.labahn@uni-rostock.de)



## Vom Tage.

Im niederösterreichischen Landtage wurde heute wieder über die Erhöhung der Spitalsgebühren nicht referirt, obwohl die festgesetzte Frist um vier volle Tage überschritten ist. Die Subcomités bewähren sich.

*Die Automatische Texterkennung liefert eine Transkription für jede Textzeile*

Die automatische Texterkennung (Automated Text Recognition, ATR) hat die Aufgabe, die in Bildern enthaltenen Textinhalte automatisch zu transkribieren. In diesem grundlegenden Schritt des Workflows werden die Textdaten extrahiert, welche in darauf folgenden Arbeitsschritten bei der Verarbeitung natürlicher Sprachelemente Anwendung finden.

In unserem Fall wenden wir Modelle auf segmentierte Textzeilenbilder an, die das Ergebnis eines Layoutanalyse-Algorithmus sind. Diese Modelle umfassen tiefe künstliche neuronale Netze, die für Sequenz-Labeling-Aufgaben entwickelt wurden. Zusätzlich enthalten die Ausgaben solcher ATR-Modelle Schätzungen für Wahrscheinlichkeiten der einzelnen Zeichen, die zur Suche in der Transkription mittels Keyword-Spotting verwendet werden können.



A Digital Investigator for  
Historical Newspapers



In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu



# NewsEye Demonstrator

## Kontakt:

Antoine Doucet

✉ antoine.doucet@univ-lr.fr



Der NewsEye-Demonstrator (<https://platform.newseye.eu>) stellt die verschiedenen Ergebnisse des Projektes vor. Seine frei zugängliche Webseite ermöglicht es den Benutzern\*innen, Zeitungssammlungen in verschiedenen Sprachen zu erkunden. Diese Zeitungen, aus dem Zeitraum von 1850 bis 1950 stammend, wurden von den

drei Nationalbibliotheken des Projektes aus Wien, Paris und Helsinki bereitgestellt. Nach der Suche in den Sammlungen können die Benutzer\*innen persönliche Sammlungen anlegen, um für sie interessante Dokumente zu verwalten. Diese können anschließend in verschiedenen Formaten exportiert oder direkt im Demonstrator analysiert werden.



A Digital Investigator for  
Historical Newspapers

In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu



## Semantische Textan- reicherung von OCR-Inhalten

### Kontakt:

Antoine Doucet

✉ antoine.doucet@univ-lr.fr

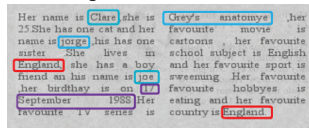
In der semantischen Textanreicherung werden Dokumente analysiert und anschließend semantische Metadaten zu deren Inhalten hinzugefügt. Bei der Vielzahl an semantischen Metadaten legen wir den Fokus auf benannte Entitäten (named entities). Genauer gesagt, wollen wir diese Entitäten in Dokumenten erfassen und zu einer Wissensdatenbank zusammenführen. Zusätzlich werden diese Entitäten mit der Haltung des Textes, in dem sie erwähnt werden, in Verbindung gebracht.

Je nach Zustand der Dokumente, bedingt durch Alterung, schlechte Lagerbedingungen und/oder die schlechte Qualität der verwendeten Druckermaterialien, kann die OCR jedoch zahlreiche Fehler aufweisen. Diese Fehler reduzieren die Leistung aller nachfolgenden Verfahren der natürlichen Sprachverarbeitung (z. B. NER, NEL und Stance Detection).

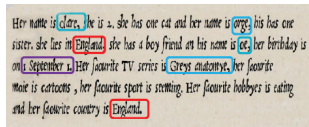
Als Lösung dieser Probleme werden in NewsEye Ansätze auf der Grundlage neuester Techniken (Neuronale Netze und Deep Learning) entworfen, die robust gegenüber OCR-Problemen

und sprachunabhängig sind. Die Ansätze von NewsEye erreichten bei der CLEF HIPE 2020 Competition zu NER- und NEL-Aufgaben in Englisch, Französisch und Deutsch von insgesamt 13 Teilnehmern in 50 von 52 Bestenlisten den ersten Platz.

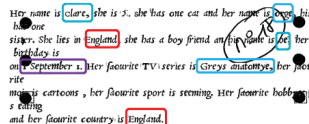
### Angemessene Qualität



### Mittlere Qualität



### Schlechte Qualität





A Digital Investigator for  
Historical Newspapers

In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

## Case Study: Gender

## Frauen (in Hosen)

**Kontakt:**

Die Case Study zu Frauen und Gender untersucht die Entwicklung der Repräsentation von Frauen in den Nachrichten zwischen 1850 und 1950. Wir interessieren uns für die Stellung der Frau in der Gesellschaft (das Frauenwahlrecht,



aber auch das Recht auf das Tragen von Hosen) sowie für jene Frauen, welche die journalistische Sphäre betraten: Journalistinnen und Schriftstellerinnen. Welchen Platz nahmen sie in den Zeitungen ein? Welche Rolle spielten die Zeitungen im Kampf um Emanzipation? Digitale Tools (z. B. Topic Modelling und Frequenzanalysen) sind in diesem Zusammenhang sehr hilfreich, um versteckte Entwicklungen und neue Trends zu entdecken.



A Digital Investigator for  
Historical Newspapers

In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Case Study: Medien & Journalismus

Die Interaktion  
zwischen Redaktion  
und Leserschaft

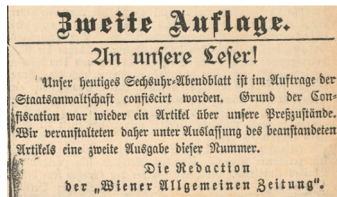
### Kontakt:

Eva Pfanzelter

✉ Eva.Pfanzelter@uibk.ac.at

## An die künftigen Abonnenten.

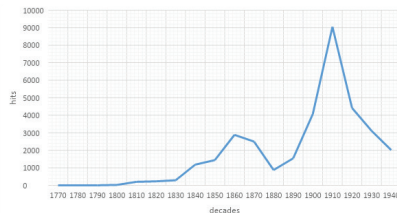
Hast du nie ein Schiff vom Stapel laufen sehen, lieber Leser? Es ist ein interessanter Anblick, und gar manche Betrachtungen und Gedanken, heiz



Wiener Allgemeine Zeitung, 13. November  
1880, Seite 1, Abendzeitung

Im „langen“ 19. Jahrhundert wurden Zeitungen zum Massenmedium schlechthin. Gleichzeitig entstand dadurch ein neuer öffentlicher Raum, in dem die Verhandlungen des modernen Nationalstaates eine wichtige Rolle spielten. Die gängige Wahrnehmung, dass es sich bei der Nachrichten- und

Meinungsverbreitung der Presse in dieser Zeit um einen einseitigen, linearen Prozess handelt, in dem die Information von der Redaktion zur Leserschaft getragen wird, ist jedoch nicht zur Gänze korrekt. Es gab unterschiedliche Möglichkeiten der Interaktion und zweiseitigen Kommunikation. In den Blick genommen wird daher in dieser Fallstudie diese Interaktion.



Absolute Häufigkeit der Phrase „An die Redaktion“  
nach Jahrzehnt



A Digital Investigator for  
Historical Newspapers



In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

# NEWS E E

## Case Study: Nationalismen

Die Sprache des  
Nationalismus  
1850-1950

### Kontakt:

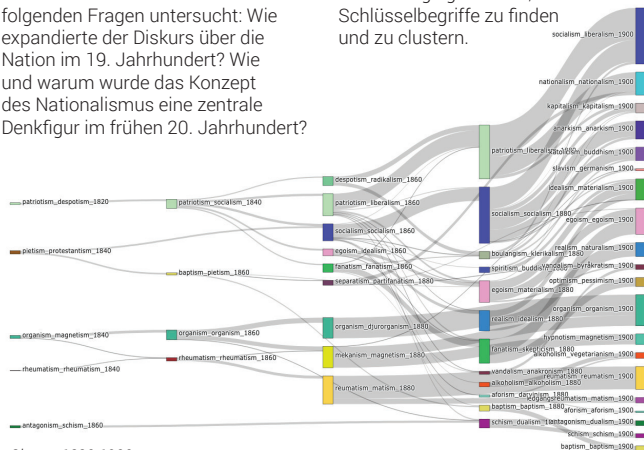
Jani Marjanen

✉ jani.marjanen@helsinki.fi

*Finska Språket, betraktadt såsom Nationalspråk.*

Die Case Study zu Nationalismen konzentriert sich auf die sich verändernde Sprache der Nation im späten 19. und frühen 20. Jahrhundert. Es werden die folgenden Fragen untersucht: Wie expandierte der Diskurs über die Nation im 19. Jahrhundert? Wie und warum wurde das Konzept des Nationalismus eine zentrale Denkfigur im frühen 20. Jahrhundert?

In der Fallstudie wird sowohl mit qualitativer Analyse von Quellentexten als auch mit Frequenzanalysen und Word Embeddings gearbeitet, um Schlüsselbegriffe zu finden und zu clustern.





A Digital Investigator for  
Historical Newspapers

In diesem Projekt arbeiten gemeinsam  
Geisteswissenschaftler\*innen,  
Informatiker\*innen und Bibliothekar\*innen  
an historischen Zeitungen  
aus drei Nationalbibliotheken  
(Österreich, Finnland, Frankreich)

 <https://www.univie.ac.at/newseye/>

 [www.newseye.eu](http://www.newseye.eu)

 @newseyeeu

## Case Study: Migration Rückkehrmigration 1850-1950

### Kontakt:

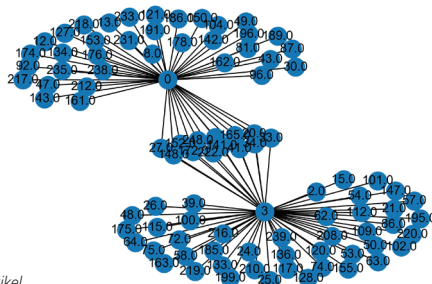
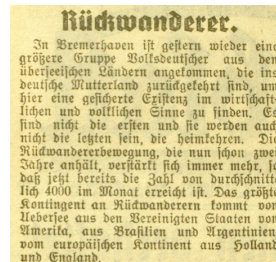
Sarah Oberbichler

✉ [sarah.oberbichler@uibk.ac.at](mailto:sarah.oberbichler@uibk.ac.at)

## Sie sind da, sie sind da!

Die Case Study zu Migration versucht, Prozesse der Remigration nach Europa zwischen 1850 und 1950 aufzuspüren, indem folgenden Fragen nachgegangen wird: Wie und in welchem Kontext berichteten österreichische Tageszeitungen über Rückkehrende und wie änderte sich diese Berichterstattung im Laufe der Zeit?

Die Fallstudie setzt digitale Methoden wie Topic Modelling und Word Embedding ein, um Subkorpora für weitere qualitative Analysen zu bilden (Diskurs- und Argumentationsanalyse).



Topic-Visualisierung für  
relevante und nicht-relevante Artikel