Grenzüberschreitendes Textmining von Historischen Zeitungen

Das impresso-Projekt zwischen Text- und Bildverarbeitung, Design und Geschichtswissenschaft

Ehrmann, Maud
maud.ehrmann@epfl.ch
École polytechnique fédérale de Lausanne - EPFL

Bunout, Estelle
estellebunout@gmail.com
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Clematide, Simon
siclemat@ifi.uzh.ch
University of Zurich

Düring, Marten
marten.during@uni.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Fickers, Andreas
andreas.fickers@uni.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Guido, Daniele
daniele.guido@uni.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Kalyakin, Roman
roman@kalyakin.com
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Kaplan, Frederic
frederic.kaplan@epfl.ch
École polytechnique fédérale de Lausanne - EPFL

Romanello, Matteo
matteo.romanello@unil.ch
École polytechnique fédérale de Lausanne - EPFL

Schroeder, Paul
hello@youtag.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Ströbel, Philip
pstroebel@cl.uzh.ch
University of Zurich

van Beek, Thijs
thijs@midasweb.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Volk, Martin
martin.volk@uzh.ch
University of Zurich

Wieneke, Lars
lars.wieneke@uni.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

impresso. Media Monitoring of the Past ist ein interdisziplinäres Forschungsprojekt, in dem Wissenschaftler und Wissenschaftlerinnen aus der Computerlinguistik, Design und den Geschichtswissenschaften an der Anreicherung eines Korpus aus schweizerischen und luxemburgischen Zeitungen arbeiten. Ziel ist es, die Qualität von Textmining-Werkzeugen für historische Zeitungstexte zu verbessern, letztere mit zusätzlichen Informationen anzureichern und schließlich mit Hilfe einer neu entwickelten Benutzeroberfläche in den historischen Forschungsprozess zu integrieren.

impresso adressiert die Herausforderungen, die große Sammlungen von digitalisierten und mit Daten angereicherten Zeitungen mit sich bringen. Diese lassen sich in fünf Kategorien zusammenfassen:

1. Dokumenten-Silos: Portale für digitalisierte Zeitungen bieten aufgrund rechtlicher Restriktionen und digitalisierungspolitischer Zwänge zwangsläufig unvollständige, nicht repräsentative Sammlungen, die einer automatisierten Verarbeitung in unterschiedlicher Qualität unterzogen wurden.

2. Große Mengen, großes Wirrwarr: Zeitungsdigitalisate sind durch Unvollständigkeit, Inkonsistenzen und Duplikate gekennzeichnet.

3. Textrauschen: Unvollkommene OCR, fehlerhafte Artikelsegmentierung und das Fehlen geeigneter linguistischer Ressourcen beeinträchtigen die Robustheit von Bild- und Text-Mining-Algorithmen erheblich.

4. Generosity (Whitelaw 2015): Suche und Auffinden relevanter Inhalte in solch großen und heterogenen Korpora.

5. Transparenz: Kritische Beurteilung der inhärenten Verzerrungen in digitalisierten Quellen und den daraus extrahierten Daten.

Parallel zu *impresso*, haben in jüngster Zeit eine Reihe von ähnlich gelagerten Forschungsprojekten computergestützte Methoden für die Analyse digitalisierter historischer Zeitungen angewandt (Ridge et al. 2019). In diesem Beitrag präsentieren wir unseren Ansatz, den obigen Herausforderungen mit Hilfe eines interdisziplinären Teams und eines Co-Design-Ansatzes gerecht zu werden (Allen and Sieczkiewicz 2010; Atanassova 2014;

Hechl et al. 2021; Ehrmann 2019). Im Folgenden geben wir einen Überblick über die wesentlichen Schritte der Dokumenten- und Textverarbeitung und deren Repräsentation innerhalb des Interfaces. Für jeden Schritt werden wir dessen Mehrwert, überwundene Schwierigkeiten und zukünftige Zielsetzungen darlegen. Ein Beispiel soll aber zunächst den Mehrwert des multilingualen, schweizerisch-luxemburgischen Korpus, der semantischen Anreicherungen und der Benutzeroberfläche illustrieren: Eine einfache Stichwortsuche nach Artikeln, die über die Schlacht um Arnheim seit dem Jahr 1944 berichten stößt in diesem Korpus schnell an ihre Grenzen: Die Stichwörter “arnheim” (deutsche Schreibweise) oder “arnhem” (niederländisch, englisch, französisch) liefern eine Vielzahl von irrelevanten Treffern verursacht durch den dortigen Fußball-Club Vitesse Arnhem, Werbeanzeigen und OCR-Fehler. Im Vergleich zu anderen Benutzeroberflächen für historische Zeitungen kann eine solche Suche in der *impresso*-App erheblich präziser formuliert und erweitert werden. Im Folgenden werden die wesentlichen Arbeitsschritte (Siehe auch Abb. 1) und Komponenten vorgestellt und, wo sinnvoll, durch das obige Beispiel illustriert.

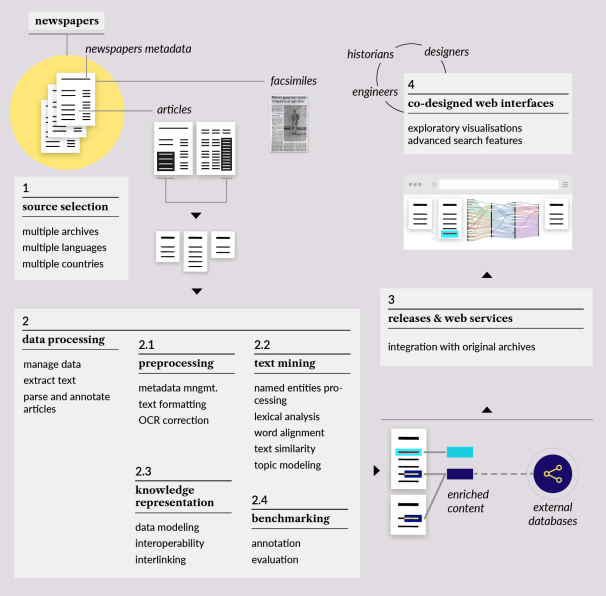

Abb. 1: Überblick über wesentlichen Arbeitsschritte des *impresso*-Projekts

Datenerfassung und Vorverarbeitung

Digitalisierte Zeitungen sind in “Silos” über viele Institutionen mit unterschiedlichen Zugriffsrichtlinien und Einrichtungen verstreut. Aus rechtlicher und administrativer Perspektive werden wir auf unsere Strategien zur Akquise und Inventarisierung digitaler Zeitungssammlungen eingehen und über unsere Vereinbarungen mit Datenanbietern berichten. Aus technischer Perspektive werden wir die Workflows skizzieren, die für den Umgang mit heterogenen Bild- und OCR-Formaten (Optical Character Recognition) entwickelt wurden, ebenso wie unsere Bemühungen um deren Standardisierung.

Digitale Dokumentenverarbeitung

Unser Ausgangspunkt sind – im Idealfall – Texte und Textblöcke, wie sie von OCR und OLR (Optical Layout Recognition) ausgegeben werden. Aufgrund der bereits erwähnten Heterogenität sind Texte und Layoutelemente jedoch meist “verrauscht” und müssen sorgfältig bewertet, korrigiert und zuweilen mittels OCR neu erfasst werden. Während des Projekts haben wir deshalb eine mehrsprachige OCR-Qualitätsbeurteilung, HTR-basierte Systeme für die Erkennung von Frakturschrift (Ströbel 2019) und die semantische Segmentierung von Zeitungen unter Verwendung textueller und visueller Merkmale (Barman 2019) getestet.

Lexikalische Verarbeitung

Nach der Texterkennung, bestand der nächste Schritt im Aufbau des impresso-Korpus in der Anwendung einer Reihe von linguistischen Vorarbeiten, einschließlich Sprachidentifikation, Tokenisierung, Normalisierung der historischen Rechtschreibung und Lemmatisierung. Historischer Sprachwandel und die Mehrsprachigkeit unseres Zeitungskorpus (französisch, deutsch, luxemburgisch) haben diese Aufgaben verkompliziert. Als ergänzende Ressourcen sind diese Vorarbeiten aber für nachfolgende Textverarbeitungsaufgaben nützlich. Desweiteren haben wir verteilte Repräsentationen von Wörtern (word embeddings) für jede Sprache im Korpus berechnet. N-Gram-Visualisierungen spiegeln die Veränderungen im Wortgebrauch in unserem Korpus im Laufe der Zeit wider; korpusspezifische word embeddings ermöglichen es dem Benutzer, verwandte Wörter zu finden und auch über verschiedene Sprachen hinweg zu vergleichen. Innerhalb des Interfaces dienen word embeddings zur Anzeige von Vorschlägen für die Schlagwortsuche und für N-gram-Analysen, indem sie synonyme oder verwandte Begriffe anzeigen (“arnheim”, “arnhem”), aber auch auf historische Schreibvarianten und häufige OCR-Fehler (“arnehm”) hinweisen. Die Benutzeroberfläche ermöglicht ebenfalls den Vergleich mehrerer N-grams in Kombination mit Suchanfragen oder innerhalb einzelner Artikel-Sammlungen, beispielsweise um die Präsenz der Schlacht von Arnheim mit jener von anderen Schlachten zu vergleichen.

Benannte Entitäten

Benannte Einheiten wie Namen von Personen, Orten und Organisationen liegen der Semantik von Texten zugrunde und sind von entscheidender Bedeutung bei deren Interpretation. Ihre automatische Erkennung und Disambiguierung unterstützt das information retrieval und die Exploration großer Textsammlungen erheblich. Sie zeigen die wechselnden Kontexte auf, in denen Personen, Institutionen und Orte über Sprachen, Zeit und Zeitungen hinweg erscheinen. Durch die Verknüpfung werden beispielsweise die Entitäten “Michail Gorbatschow” und “Mikhail Gorbachev” derselben Person zugeordnet und durch kontextualisierende Informationen wie Lebensdaten angereichert. Innerhalb des Interfaces kann die Verteilung der Entitäten innerhalb des Korpus, auch über Sprachgrenzen hinweg verfolgt werden.

Topic Modelling

Wir ermitteln sprachspezifisch, welche "Themen" in unserem Zeitungskorpus vorkommen, um daraus eine Thematik für den Benutzer abzuleiten. Zu diesem Zweck werden mehrere topic models (über das gesamte Korpus, pro Zeitung, pro Zeitraum) berechnet und die Themen den Zeitungsartikeln zugeordnet. Topics erlauben es, überwältigende Ergebnismengen zu reduzieren, bestimmte Themenaspekte wie “Sport”, “Militär” oder “Kunst” ein- bzw. auszuschließen. Im obigen Beispiel würden also sportbasierte Themen aus- und militär-basierte Themen eingeschlossen. Innerhalb des Interfaces visualisiert eine Graph-Visualisierung Überschneidungen zwischen Topics. Knoten repräsentieren Topics, eine Kante zwischen zwei Knoten ist gewichtet und basiert auf der Zahl der Artikel, in denen beide Topics erkannt wurden.

Text reuse

Text reuse erkennt und vergleicht ähnliche Textpassagen und liefert Cluster von wiederverwendeten Text-Passagen, die sich in großen Sammlungen von Dokumenten auffinden lassen. Innerhalb des impresso-Projekts haben wir passim (Romanello 2018; Smith et al. 2014) verwendet. Das Interface zeigt Text reuse Passagen innerhalb von Artikeln an und eine eigene Komponente erlaubt die Stichwortsuche in allen Text reuse Clustern mit diversen Filter-Möglichkeiten um beispielsweise die Zirkulation von Agenturberichten über die Schlacht von Arnheim in der Presselandschaft zu rekonstruieren.

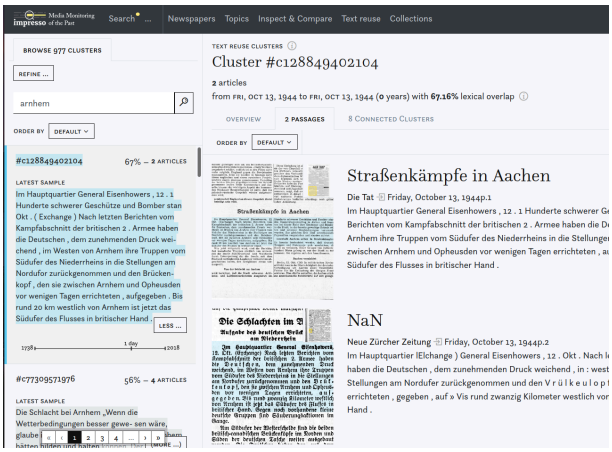

Abb. 2: Text Reuse Explorer

Bildähnlichkeits-Suche

Eine auf (Seguin 2018) basierende visuelle Suchmaschine ergänzt die Textsuche und erlaubt es, Kopien und einander ähnliche visuelle Elemente, wie z.B. Anzeigen, Fotos, Zeichnungen und Karten zu suchen. Die Bildähnlichkeitssuche erlaubt drei Formen der Interaktion: ausgehend von einem Bild können ähnliche Bilder gefunden werden, Bilder können über eine Stichwortsuche im umgebenden Text gefunden werden, externe Bilder können hochgeladen werden um zu prüfen, ob ähnliche Bilder im Korpus vorhanden sind.

Systemarchitektur

Text- und Bildverarbeitungskomponenten müssen in eine modulare Systemarchitektur integriert werden, die auch eine API, ein middle layer und ein front end umfasst. Wir haben eine technische Dokumentation veröffentlicht, die Informationen zu allen Schritten in der Aufbereitung und Anreicherung unseres Korpus und einer API enthält.

Die Benutzeroberfläche wurde mit dem Ziel entwickelt, quellen- und datenkritisches Textmining für die breite Masse von historisch arbeitenden WissenschaftlerInnen anzubieten und neue Arbeitsabläufe für die Exploration historischer Zeitungen zu ermöglichen. Um eine hohe Anschlussfähigkeit an gängige Forschungspraxis zu gewährleisten, haben wir den Schwerpunkt auf das Auffinden von relevanten Inhalten gelegt. Inspiriert wurde die Gestaltung der Benutzeroberfläche durch die Prinzipien der Generosity, einem sinnbildlichen sich-öffnen des Korpus und der Transparenz, die eine informierte Nutzung und kritische Einordnung der gemachten Beobachtungen ermöglichen soll. Weiter angeregt durch die Gestaltung durch das in mehreren Workshops gesammelte Benutzerfeedback und durch das übergreifende Ziel, nahtlos zwischen close und distant reading Perspektiven zu wechseln. Um eine möglichst breite Anwendbarkeit und unterschiedlichste Workflows zu gewährleisten, wurde die Benutzeroberfläche so gestaltet, dass sie NutzerInnen im Rahmen des technisch möglichen die freie Kombination aller obenstehenden semantischen Anreicherungen und den auf ihnen basierenden Komponenten erlaubt. Dies wird ermöglicht durch eine Reihe von Komponenten, die innerhalb der gesamten Benutzeroberfläche verfügbar sind.

Search

Suchoperationen gehören zu den am häufigsten genutzten Interaktionen mit digitalisierten Quellen. Innerhalb der Benutzeroberfläche wurden deshalb Merkmale der klassischen “erweiterten” Suche mit semantischen Anreicherungen verknüpft um im Sinne der Generosity unterschiedliche Zugänge in das Material zu ermöglichen. Neben klassischen UND/ODER/NICHT Interaktionen, schlägt die Suche verwandte oder synonyme Suchbegriffe auf Basis des Korpus vor ebenso wie relevante Entitäten und Topics. In Kombination mit Filtern, die sowohl auf Bestandsmetadaten und semantischen Anreicherungen basieren, ermöglicht es die Suche in einem iterativen Prozess hochkomplexe Suchanfragen zu formulieren.

Abb. 3: Suche mit Hilfe von word embeddings, topics, Spracherkennung und Bestandsmetadaten

Collections

Ermöglichen den manuellen oder Suchabfragen-basierten Aufbau themenspezifischer Sammlungen mit bis zu 10.000 Artikeln, die als eigene Objekte ebenfalls durchsucht und gefiltert werden können.

Inspect&Compare

Inspect&Compare erlaubt den Vergleich von Suchanfragen und Sammlungen hinsichtlich ihrer Überlappungen und Differenzen mit Hilfe von Zeitleisten und Balkendiagrammen. Beispielsweise ließen sich eine Sammlung von Artikeln zu Arnheim mit einer Sammlung über die Schlacht von El-Alamein hinsichtlich ihrer Präsenz in der Medienberichterstattung vergleichen. Ebenso eignet sich die Komponente für den iterative Aufbau von Sammlungen durch experimentelle Variationen von Suchanfragen und dem Vergleich ihrer Ergebnisse. Figure X zeigt beispielsweise an, dass die angelegte Sammlung in (A) 182 Artikel mit dem Begriff „vitesse arnhem“ enthält.

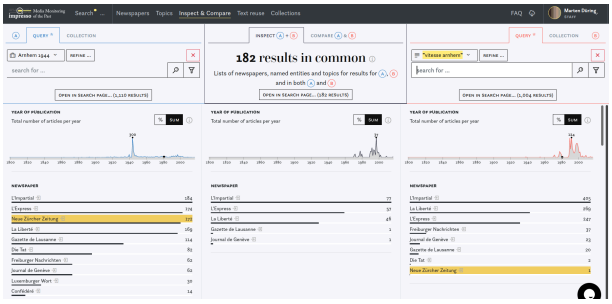

Abb. 4: Inspect&Compare, Vergleich einer Sammlung mit einer Suchabfrage

Überblicks-Fenster

Zu jedem Objekt innerhalb der Benutzeroberfläche, wie z.B. genannten Entitäten, Topics oder Zeitungen erlaubt ein korrespondierendes Überblicks-Fenster deren Präsenz innerhalb der laufenden Suche bzw. des Korpus zu prüfen und weitere Informationen, wie z.B. Metadaten abzurufen.

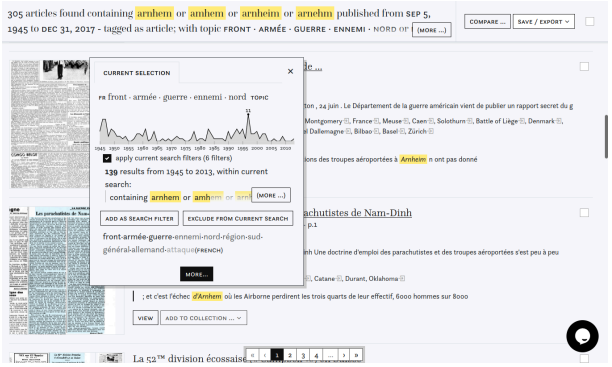

Abb. 5: Überblicks-Fenser eines topics

Export

Artikelsammlungen können mit allen semantischen Anreicherungen für die weitere Analyse außerhalb des Interfaces als .csv Dateien exportiert werden, je nach Rechtssituation auch inklusive des Artikel-Volltextes.

Artikelempfehlungen

Such- und Filteroperationen dienen dazu, relevante Inhalte innerhalb des Korpus zu identifizieren und neue Facetten auf das Quellenmaterial zu erhalten. Ein automatisiertes Empfehlungssystem hilft dabei, relevante Inhalte zu finden, die außerhalb des Suchbereichs der NutzerInnen liegen. Artikelempfehlungen basieren auf topics, zeitlicher Distanz, benannten Entitäten und text reuse, die voneinander unabhängig gewichtet werden können. Dies könnten Artikel sein, die zwar die Schlacht von Arnheim thematisieren und ähnliche benannte Entitäten erwähnen, aber nicht einem militär-basierten Topic zugeordnet wurden und deshalb aus den vorigen Ergebnislisten ausgeschlossen wurden.

Korpus-Übersicht

Visualisierung des Gesamtbestandes inklusive aller Bibliotheks-Metadaten, die auch Lücken, Verzerrungen im Korpus einschließt.

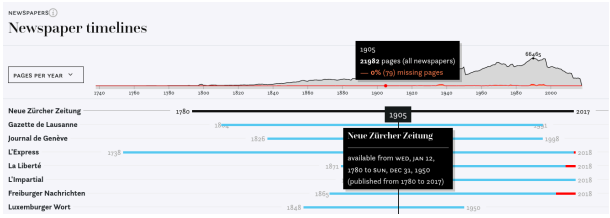

Abb. 6: Korpus-Übersicht

Lehr-/Lernmaterialien

Die Gestaltung der Benutzeroberfläche hat sich an den durchschnittlichen technischen Kompetenzen der historisch arbeitenden GeisteswissenschaftlerInnen orientiert, mit dem Ziel, NutzerInnen zu fordern ohne sie zu überfordern. Gleichzeitig, dem Anspruch der Transparenz folgend, sollten möglichst alle relevanten Methoden und Entscheidungen im Einreichungsprozess erklärt und dokumentiert werden. Diesen Zwecke erfüllen eine Reihe von FAQs, Blogartikeln, Tutorials und Videos, die das Projekt und die Funktionalitäten der Benutzeroberfläche dokumentieren.

Die Kombination aus semantischer Anreicherung, Design und historischen Forschungsinteressen hat - wie wir hier illustriert haben – zu einer Vielzahl von neuen Interaktionsmöglichkeiten mit historischen Zeitungen geführt. In Kombination tragen diese dazu bei, relevante Inhalte besser zu finden, close und distant reading zu integrieren und Vergleichsperspektiven einzunehmen.

Bibliographie

Allen, Robert, and Robert Sieczkiewicz. 2010. “How Historians Use Historical Newspapers.” *Proceedings of the American Society for Information Science and Technology* 47. https://doi.org/10.1002/meet.14504701131.

Atanassova, Rossitza. 2014. “Improving the Discovery of European Historic Newspapers.” In *IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge*. Lyon, France. http://library.ifla.org/1038/.

Barman, Raphaël. 2019. “Historical Newspaper Semantic Segmentation Using Visual and Textual Features.”

Ehrmann, Maud. 2019. “Historical Newspaper User Interfaces: A Review.” In *IFLA 85* th . Athens, Greece: IFLA. http://library.ifla.org/2578/.

Hechl, Stefan, Pierre-Carl Langlais, Jani Marjanen, Sarah Oberbichler, and Eva Pfanzelter. 2021. “Digital Interfaces of Historical Newspapers: Opportunities, Restrictions and Recommendations.” *Journal of Data Mining & Digital Humanities* HistoInformatics (January). https://doi.org/10.46298/jdmdh.6121.

Ridge, Mia, Giovanni Colavizza, Laurel Brake, Maud Ehrmann, Jean-Phillipe Moreux, and Andrew Prescott, eds. 2019. “The Past, Present and Future of Digital Scholarship with Newspaper Collections.” In *DH 2019 Book of Abstracts*.

Romanello, Matteo. 2018. “Detecting Text Reuse in Newspapers Data with Passim.” Presented at the Hacking the News Workshop in conjunction with DHN 2018, Helsinki. http://dig-humnord.eu/.

Seguin, Benoît Laurent Auguste. 2018. “Making Large Art Historical Photo Archives Searchable.” Lausanne: EPFL.

Smith, David A., Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. “Detecting and Modeling Local Text Reuse.” In *IEEE/ACM Joint Conference on Digital Libraries*, 183–92. https://doi.org/10.1109/JCDL.2014.6970166.

Ströbel, Phillip. 2019. “Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images.” Presented at the *Digital Humanities 2019: Complexities, Utrecht, December 7*. https://www.conftool.pro/dh2019/index.php?page=brows%20eSessions&path=adminSessions&print=export&ismobile=%20false&form_session=481&presentations=show.

Whitelaw, Mitchell. 2015. “Generous Interfaces for Digital Cultural Collections” 9 (1). http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html.