Phillip Ströbel
Universität Zürich, Institut für Computerlinguistik

Computerlinguistische Methoden für bessere Zugänglichkeit von historischen Zeitungsberichten: Die NZZ im Wandel der Zeit

Im digitalen Zeitalter werden Zeitungen auf Papier obsolet. Wegen sinkender Print-Auflagen setzen Medienhäuser auf den Online-Markt. Dies war bis vor 30 Jahren noch undenkbar, denn welche Medien außer gedruckte Zeitungen, Radio und Fernsehen informierten täglich, umfassend und kompetent über die Geschehnisse des Alltags? Durch die kontinuierliche Aufzeichnung vielfältiger Informationen und Meinungsäußerungen entwickeln sich die Archive der Printmedien zu einer immer wichtigeren Ressource für die Geschichtsforschung. Genau hier setzt das Ende 2017 gestartete impresso-Projekt an. Denn obwohl wir uns mitten in der Digitalisierung befinden, liegen viele Zeitungsbestände noch nicht in geeigneter Form vor. Qualitätsprobleme der OCR, der automatischen Layoutanalyse und der Artikelerkennung erschweren es, effizient auf deren Inhalte zuzugreifen. So sehen sich Forschende, die Digitalisate nutzen wollen, noch immer mit mühsamer Handarbeit in uneinheitlichen Suchportalen konfrontiert.

Das interdisziplinäre impresso-Team, bestehend aus Forschungsgruppen aus Luxemburg, Lausanne und Zürich, entwickelt ein einheitliches Suchportal und Text-Mining-Werkzeuge. Diese erleichtern die Forschungsarbeit und ermöglichen, grosse Zeitungsbestände mühelos zu durchforsten. Im Rahmen von impresso definieren Forschende der Geschichte und der Sprachtechnologie gemeinsam (Codesign-Prinzip) die relevanten Funktionalitäten. Ein oft auftauchendes Bedürfnis ist eine facettierte Ad-hoc-Suche, welche eine Verfeinerung oder Filterung der Treffer aufgrund von Themenbereichen wie „Krieg“ oder „Politik“ erlaubt („Topic Modeling“). Die 53.000 Ausgaben der Neuen Zürcher Zeitung (NZZ) von 1780 bis 1950 mit über 560.000 Seiten sind Teil der impresso-Kollektion. Anhand der Daten der NZZ möchten wir exemplarisch Probleme der Qualität und Zugänglichkeit der Digitalisate aufzeigen und auf Möglichkeiten der verbesserten automatischen Erschließung mit sprachtechnologischen Werkzeugen eingehen.