Ein Klick in die Vergangenheit

Historische Zeitschriften sind ein flüchtiges Gut. Mit Hilfe der Digitalisierung können sie nicht nur konserviert, sondern auch für die Nutzer viel einfacher zugänglich gemacht werden. An der Entwicklung zweier Nutzerplattformen - eine für Historiker und historisch Interessierte, eine für Sprachwissenschaftler - arbeiten derzeit Landesbibliothek Dr. F. Teßmann und EURAC. Eine mitunter kniffelige Aufgabe.

von **Sigrid Hechensteiner**

Am 26. Juli 1916 ruft der Tiroler Volksbote erneut zur Metallablieferung auf: Badewannen aus Kupfer (auch verzinnt oder mit anderen Metallen überzogen). Obsteinsiedekessel soweit sie nicht in fabriksmäßigen Betrieben verwendet werden... Händler und Handwerker haben, wenn sie ein Drittel schon abgeliefert haben, ein weiteres Drittel zu liefern, heißt es da. Auf derselben Seite steht unter dem Titel Besitzwechsel: Thomas Langhofer, Viehhändler in St. Johann i.T., hat von Simon Lanzinger und Alois Huber das Gut Hinterbautzenberg in Oberndorf um 10.000 Kronen käuflich erworben.

Tageszeitungen, vor allem historische, liefern einen endlosen Kosmos an Informationen. Für Historiker, Chronisten, Ahnenforscher. Doch die Recherche war bis vor kurzem zeitaufwändig. Das Ergebnis oft ungewiss. Heute macht der schnelle Zugriff auf Informationen historisches Zeitungsmaterial auch für Laien attraktiv. Eine simple Begriff-Eingabe auf der Suchleiste reicht und schon spuckt etwa die Plattform Teßmann Digital Zeitungsausschnitte aus, die bis 1813 zurückreichen. „Rund 90% der Plattformnutzer suchen über Namen-, Datums- oder Ortseingaben nach einem Stück Vergangenheit", erzählt Johannes Andresen, Direktor der Landesbibliothek Dr. F. Teßmann. Im Monat verzeichnet Teßmann Digital rund 1,6 Millionen Klicks. Tendenz steigend. Nach dem Relaunch im November 2014 waren es schon 36 Prozent mehr Zugriffe als im Vormonat. „Im Schnitt verbringen unsere Nutzer 16 Minuten auf unserem digitalen Zeitungsarchiv", so Johannes Andresen. Das sei beachtlich, wenn laut Chartbeat-Studie über die Hälfte der Leser maximal 15 Sekunden auf einer Internetseite verweilen.

Die Maschine kann nun mit neuen Dokumenten gefüttert werden, und sie erledigt Wortkorrekturen von alleine. Fünf Seiten pro Sekunde schafft sie.

Das Internet ist zum Forschungsinstrument geworden und das nicht nur für Spezialisten.
Doch die Digitalisierung und nutzergerechte Aufbereitung alter Bestände stellt Bibliothekare und Programmierer auch immer wieder vor neue Herausforderungen. „Ganz so einfach ist es nicht", sagt EURAC-Computerlinguist Michel Généreux. „Bevor wir mit der eigentlichen Programmierarbeit beginnen, müssen die eingescannten Zeitungstexte durch Texterkennungsprogramme laufen, und meist treten da schon die ersten Probleme auf."

Der gebürtige Kanadier weiß, wovon er spricht. Seit Monaten ist er damit beschäftigt die Qualität von 100.000 digitalisierten Zeitungsseiten zu verbessern. Bei dem Material handelt es sich um elf deutschsprachige, Südtiroler Tageszeitungen aus dem Zeitraum 1910 bis 1920. Sie bilden die Grundlage für das Projekt OPATCH (*Open Platform for Access to and Analysis of Textual Documents of Cultural Heritage*), an dem Teßmann und EURAC gemeinsam arbeiten. Ziel ist es, aufbauend auf dem Material zwei

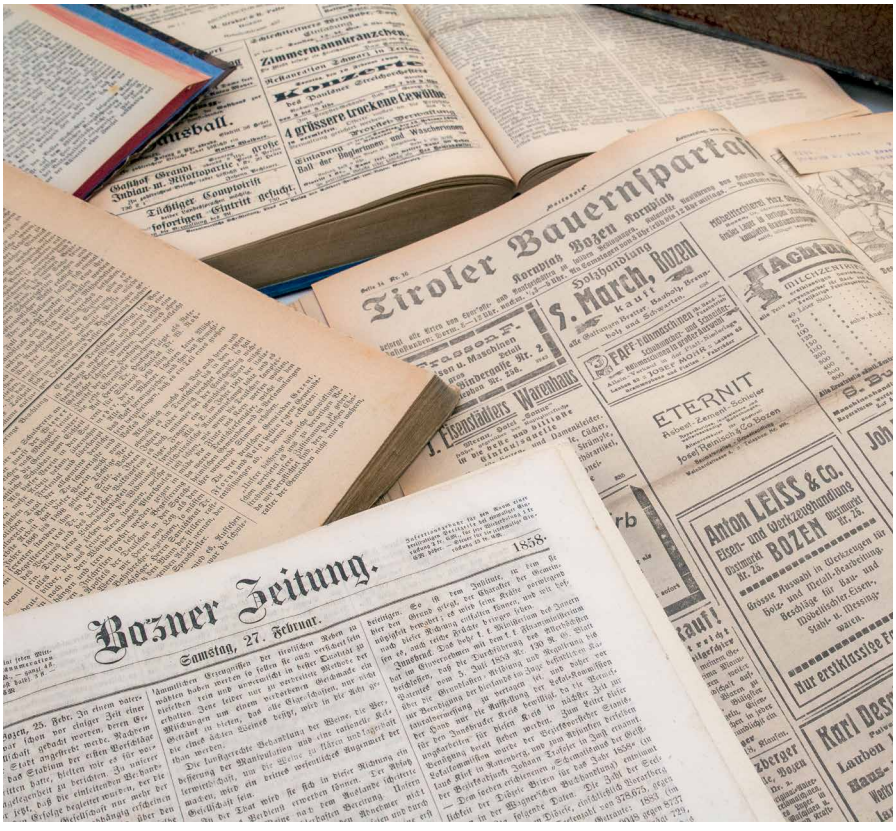
01

Nutzerplattformen zu realisieren: Die eine richtet sich an Historiker und historisch Interessierte, die andere an Linguisten.

Warum die Vorarbeiten so viel Zeit in Anspruch nahmen? „Das lag zum einen an der Schriftart, Fraktur, bei der Buchstaben häufig verwechselt werden", erklärt der Computerlinguist Généreux, „zum anderen am holzhaltigen Papier." Die Seiten sind übersäht mit gelben Flecken und zerbröseln allmählich. Der Computer tut sich schwer, Flecken oder Risse von Buchstaben zu unterscheiden.
Um alle möglichen Fehlerquellen aufzuspüren hat eine Teßmann-Mitarbeiterin zunächst zehn ausgewählte, volltexterkannte Seiten von Hand korrigiert. Diese *Gold Standard* Seiten wurden in den Computer eingespeist und mit den entsprechenden zehn digitalisierten Seiten abgeglichen. Häufige Fehlerquellen waren Verwechslungen der Buchstaben n (n) und u (u) sowie ſ (s) und f (f).
So hat Généreux Algorithmen programmiert, die alle diese Buchstaben automatisch kontrollieren und wo nötig richtigstellen. Liest der Computer etwa „Urfache", korrigiert er „Ursache". Damit der Computer weiß, dass „Ursache" als Wort existiert, muss er auf Referenzdaten zugreifen können. Im Falle von OPATCH ist es ein frei zugängliches Web Corpus fürs Deutsche, gepaart mit Texten des Projekts Gutenberg, da es sich um Sprachgebrauch des frühen 20. Jahrhunderts handelt.
„Die Maschine kann nun mit neuen Dokumenten gefüttert werden, und sie erledigt Wortkorrekturen von alleine", so Généreux. Fünf Seiten pro Sekunde schafft sie.

Für die Realisierung der Bibliotheksplattform wurden im Vorfeld potentielle Nutzer interviewt.
Diese suchen am häufigsten über die Eingabe von Namen (Eigennamen, Namen von Unternehmen oder Organisationen), Orten und Daten. Also wird in der Computerlinguistik Wert auf die so genannten *Named Entities* gelegt. Damit der Computer beispielsweise schnell lernt, Personennamen zu erkennen, wurden bei OPATCH zunächst händisch 200 Seiten an Namen (Personennamen, Orte, Organisationen) *getagged*. Bald lernt der Computer, dass es sich bei „Andreas" mit hoher Wahrscheinlichkeit um einen Vornamen handelt. Und dass einem Vornamen meist auch ein Nachname folgt. Mit Hilfe der Kontextanalyse ist er dann sogar in der Lage, die Suche nach „Andreas Hofer" nur auf den reinen Namen zu beschränken. „Andreas Hofer Straße" wird er nicht als Suchergebnis angeben.
Ebenso muss der Computer lernen, dass beispielsweise „Bozen" in den Zeitungstexten auch oft „Stadt an der Etsch" genannt wird, oder aber „Augsburg" auch die „Fuggerstadt" ist.
Linguisten wiederum haben ganz andere Bedürfnisse. Sie möchten etwa syntaktische Zusammenhänge verstehen oder untersuchen, wie sich unterschiedliche Wortklassen (Verben, Nomen, Adjektive, usw.) im Satzkontext verhalten. Für derartige Analysen kann es hilfreich sein, graphische Darstellungsformen, sogenannte Visualisierungen, einzusetzen. Für statistische Untersuchungen ist es praktisch, in einem Volltext alle Verben farblich visualisieren zu können. Damit gewinnt man einen schnellen Überblick über die Wahl der Verben. Die Häufigkeit mit der ein Wort vorkommt, lässt sich über die Größe repräsentieren. Je größer das Wort in der Visualisierung des Volltextes, desto häufiger wird es gebraucht.
Die Programmierarbeit an den beiden Plattformen soll 2015 abgeschlossen sein. Die Ergebnisse der Plattform für Linguisten wird in die Plattform „Korpus Südtirol" miteinfließen. Jene für die historischen Zeitungen in die Seite: digital.tessmann.it. Für die Landesbibliothek erwartet sich Johannes Andresen dann nochmals einen beachtlichen Nutzerzuwachs. „Ich habe ja selbst schon meinen Namen in die Suchleiste eingegeben und gestaunt, was über meine Vorfahren so alles berichtet wurde."

O2



01 Das OPATCH-Team (v.l.n.r): Johannes Andresen, Verena Lyding, Christoph Moar, Katalin Szabo, Michel Généreux
02 Fleck oder Buchstabe? Der Computer muss unterscheiden lernen.

OPATCH wird über das Landesgesetz 14 durch die Autonome Provinz Bozen - Südtirol, Abteilung Bildungsförderung, Universität und Forschung finanziert.