Die Digitale Rätoromanische Chrestomathie

Jürgen Rolshoven
unter Mitarbeit von Florentin Lutz,
Claes Neuefeind und Fabian Steeg

Hans Dieter Bork zum 12.7.2012

1. Einführung

Der vorliegende Beitrag hat die digitale Tiefenerschließung der Rätoromanischen Chrestomathie von Caspar Decurtins zum Gegenstand. Die Rätoromanische Chrestomathie, 1896–1919 in der Zeitschrift "Romanische Forschungen" (Erlangen) erschienen, ist die bis heute wichtigste Textsammlung für das Bündnerromanische und eine für Sprach-, Literatur- und Kulturwissenschaften unübertroffene Quelle. Mit der Tiefenerschließung werden die Volltexte der Chrestomathie und darüber hinaus ein rätoromanisches Textkorpus als Grundlage korpuslinguistischer und philologischer Forschungen der Öffentlichkeit über das Internet frei zugänglich bereitgestellt. Damit verbunden ist auch ein quelloffenes Korrektur-, Erschließungs- und Anreicherungsverfahren. Es werden automatische und interaktive Korrekturen und Anreicherung kombiniert. Die interaktive Korrektur und Anreicherung setzt Wiki-Prinzipien um und bindet in Zusammenarbeit mit den Rätoromanen in der Schweiz und deren Organisationen die Sprachgemeinschaft und die am Bündnerromanischen Interessierten ein. Die im Vorhaben eingesetzten Techniken sollen in der Folge auf weitere Textsammlungen des Bündnerromanischen und anderer Sprachen angewandt werden. Das Projekt hat somit prototypischen Charakter für Vorhaben, die der Erschließung spezialisierter Textsammlungen gewidmet sind, die vorrangig die Dokumentation und

Ladinia XXXVI (2012), 119–151
ISSN 1124–1004;

Bewahrung kleinerer, auch bedrohter Sprachen zum Gegenstand haben, und die die Sprecher dieser Sprachen in Textkorrektur, -anreicherung und -nutzung einbinden. Dadurch gewinnt die Auseinandersetzung mit der eigenen Sprache und Kultur eine neue quantitative und qualitative Dimension.

Die Digitale Rätoromanische Chrestomathie ist ein über den Zeitraum von November 2009 bis Oktober 2011 von der Deutschen Forschungsgemeinschaft gefördertes Projekt der Sprachlichen Informationsverarbeitung, Institut für Linguistik, Universität zu Köln (J. Rolshoven) und der Universitäts- und Stadtbibliothek Köln (Wolfgang Schmitz).

2. Caspar Decurtins

Die Rätoromanische Chrestomathie ist das wohl bedeutendste Werk, das C. Decurtins geschaffen und hinterlassen hat. Decurtins ist eine der großen und prägenden Persönlichkeiten der Bündnerromanen. Zu Recht trägt er den Beinamen Löwe von Trun. Decurtins wurde als Sohn eines Arztes und Landammans[1] 1855 in Trun geboren. Er besuchte die Gymnasien in Disentis und Chur und widmete sich nach 1875 in Heidelberg und München dem Studium der Geschichte, der Kunstgeschichte und des Staatsrechts. Nach seiner Promotion hängte er ein weiteres Semester in Straßburg an und trat dann 1877 im Alter von 21 Jahren in das politische Leben ein. Von 1877 bis 1883 ist er Landamman (romanisch *mistral*) in der Cadi, der Landschaft um Disentis in der oberen Surselva. In dieser Funktion trägt er wesentlich zur Restaurierung des Klosters Disentis bei, nicht zuletzt auch als Ausdruck seiner politischen Orientierung, die sich dem radikalen Kulturkampf widersetzt. Decurtins ist von 1877 bis 1904 Mitglied des Großen Rates des Kantons Graubünden, von 1881 bis 1905 Mitglied des Nationalrats in Bern. Decurtins ist als Fraktionschef einer der prominentesten Vertreter der Schweizerischen Konservativen Volkspartei; sein politisches Programm und das seiner Partei sind eine Antwort auf den Kulturkampf der Radikalliberalen und den Klassenkampf der damaligen Sozialdemokraten.[2] Sozialpolitisch setzte er sich in der Arbeiterfrage für Sonntagsruhe, für Kranken- und Unfallversicherung sowie für das Fabrikgesetz ein und zeigte dabei auch keine Berührungsängste zur politischen Linken. Die Belange der Bergbevölkerung lagen ihm am Herzen. Er beriet Papst Leo XIII., als dieser der katholischen Soziallehre in der Enzyklika *Rerum Novarum* eine Grundlage gab.

[1] Der Landamman ist das gewählte Oberhaupt einer Landsgemeinde, cf. "Landamman" im Historischen Lexikon der Schweiz: <www.hls-dhs-dss.ch/textes/d/D10256.php>.

[2] Cf. auch Wigger 1997.

Nach 1900 entfremdete sich Decurtins von seiner Partei. Er trug die Hinwendung zu den Liberalen in einem Bürgerblock nicht mit. Er wurde Professor für Kulturgeschichte an der Universität Freiburg im Üechtland, für deren Gründung er sich sehr eingesetzt hatte. Die 1889 gegründete Universität war eine Hochschule des schweizerischen Katholizismus. Decurtins starb 1916 in seiner Heimatgemeinde Trun.

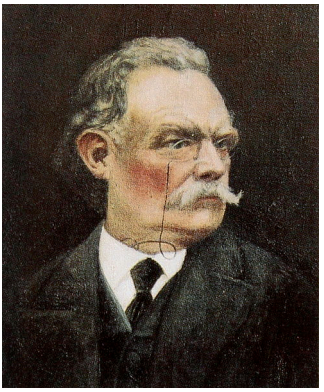


Fig. 1: Caspar Decurtins (aus: <de.wikipedia.org/wiki/Caspar_Decurtins>)

Über sein politisches – vor allem auch enormes verbandspolitisches – Wirken hinaus ist Decurtins' Name unauflöslich mit der von ihm geschaffenen Rätoromanischen Chrestomathie verbunden.[3]

3. Die Rätoromanische Chrestomathie

Aufgrund ihrer kulturwissenschaftlichen Bedeutung wird die Rätoromanische Chrestomathie als das "Herzstück der sog. Rätorom. Renaissance der Jahrhundertwende" betrachtet.[4]

Decurtins hatte ursprünglich zwei Bände geplant; im Laufe von Jahrzehnten wuchs das Werk, das zwischen 1896 (Band 1, 1. Lieferung schon 1891 in den "Romanischen Forschungen") und 1919 veröffentlicht wurde, auf 13 Bände heran. Die Textauswahl ist differenziert und reich, es sind die "verschiedenartigsten Quellengattungen vertreten, und die Texte stammen aus allen bündnerromanischen Gegenden und aus allen Jahrhunderten, in denen man diese Sprache zu Papier brachte".[5]

Die lange Entstehungsgeschichte hinterließ ihre Spuren in der Systematik und in der Zusammensetzung des Werks. Angesichts des riesigen Umfangs verwundert sicherlich nicht, dass aus heutiger Sicht editionsphilologische Schwächen erkennbar sind. Auch sind dem langen Erstellungs- und Publikationszeitraum geschul-

[3] Cf. Historisches Lexikon der Schweiz, s.v. Decurtins, Caspar: <www.hls-dhs-dss.ch/textes/d/D3565.php>.

[4] Cf. <www.hls-dhs-dss.ch/textes/d/D3565.php>; cf. auch Deplazes 1986, 5–6, Liver 1999, 82 und Riatsch 1998, 176–198.

[5] Egloff/Mathieu 1986, 7.

dete Inkonsistenzen und Perspektivwechsel zu beobachten.[6] Dies hat der Nutzer der gedruckten Fassungen zu beachten, und dies ist auch für das Potential einer digitalen Tiefenerschließung bedeutsam (cf. dazu Kap. 8, Potentiale).

Die Rätoromanische Chrestomathie wurde auch als 13-bändiger Separatdruck publiziert.[7] Nach dem Tod von Decurtins 1916 wurden die letzten beiden Bände von seiner Witwe Maria Decurtins Geronimi und von Christian Caminada, dem späteren Bischof von Chur, veröffentlicht.

Steht der Wert der Rätoromanischen Chrestomathie außer Zweifel,[8] so war ihre Rezeption in der bündnerromanischen Sprachgemeinschaft und auch in der wissenschaftlichen Fachwelt durch mancherlei Faktoren behindert. Gewiss blieben die in Deutschland herausgegebenen "Romanischen Forschungen", in denen die Chrestomathie zwischen 1896 und 1919 publiziert worden war, den Muttersprachlern wohl unbekannt oder zumindest fremd. Gleiches gilt für die Bände des Separatdrucks, die ebenfalls in Erlangen, im Verlag Junge, erschienen. Das Gesamtwerk erschließt sich in seiner auf viele (Zeitschriften-)Bände verteilten Heterogenität nur schwer. Die thematische Vielfalt, die unterschiedlichen Texte und Genres, allen Idiomen entnommen und einen Zeitraum von vier Jahrhunderten überspannend – all das verlangte der Sprachgemeinschaft viel ab. Auch fehlten Indizes und Register, die den gezielten Zugriff auf das große Werk ermöglicht hätten.

Vor diesem Hintergrund war der in den 1980er Jahren gefasste Entschluss, die Texte der Chrestomathie in einem 13-bändigen Reprint herauszugeben und durch einen Registerband zugänglich zu machen, höchst begrüßenswert. Dafür engagierten sich die "Societad Retorumantscha" unter ihrem Präsidenten Gion Deplazes und Andreas Joos und sein Octopus Verlag in Chur; sie wurden unterstützt durch zahlreiche Persönlichkeiten, Stiftungen und Kulturorganisationen Bündens und der Schweiz.[9] Peter Egloff und Jon Mathieu haben das große Verdienst, die Rätoromanische Chrestomathie den Muttersprachlern und der wissenschaftlichen Welt in sehr unterschiedlichen Perspektiven durch einen Registerband erschlossen zu haben. Zu den unterschiedlichen Erschließungsperspektiven zählen ein Personen-, ein Orts- und Sachregister, eine Klassifikation nach Inhalt und Motiven, Übersichten über räumliche und zeitliche Situierung der Texte, aber auch text- und entstehungskritische Kommentierung.

[6] Cf. op. cit., 334.

[7] Zu den Unterschieden beider Publikationen cf. op. cit., 18–19.

[8] Cf. op. cit., 337.

[9] Cf. Deplazes 1986, 5–6; Egloff/Mathieu 1986, 7.

Zu den bestehenden 13 Bänden kamen zusätzlich der Band XIV mit den Texten aus der Val Schons und der Band XV, der Registerband. Die erste 13-bändige Erlanger Auflage erschien broschiert, die zweite, zwischen 1982 und 1985 erschienene Auflage ist einheitlich gebunden.

4. Die Digitale Rätoromanische Chrestomathie

Die Rätoromanische Chrestomathie ist eine der wichtigsten Textbasen des Bündnerromanischen. Digital steht die Rätoromanische Chrestomathie bislang nur in Form von Faksimiles zur Verfügung;[10] ein elektronischer Volltextzugriff ist nicht gegeben. Eine volltexterschlossene Rätoromanische Chrestomathie ist für nahezu alle sprach- und kulturwissenschaftlichen Disziplinen von außerordentlichem Interesse; sie stimuliert lexikographisches und lexikologisches, morphologisches und syntaktisches, semantisches und textlinguistisches, literaturwissenschaftliches, volkskundliches und historisches Arbeiten. Sie ermöglicht der Sprachwissenschaft datengetriebene Untersuchungen zu Strukturen und Textsorten. Aufgrund des hohen Varietätenreichtums der Rätoromanischen Chrestomathie profitieren auch diachrone (über vier Jahrhunderte reichende) und diatopische (fünf Hauptidiome umfassende) Untersuchungen von der Volltexterschließung. Das Rätoromanische ist für die korpuslinguistische Forschung auch deshalb von besonderem Interesse, da es über die gängigen korpuslinguistischen Fragestellungen hinaus von großem Wert für Fragen von Sprachkontakt, Sprachverwandtschaft und Sprachwandel ist.[11] Der Varitätenreichtum findet nicht zuletzt auch in den unterschiedlichen Verschriftungs- und Orthographietraditionen des Bündnerromanischen einen höchst vielfältigen Ausdruck. Mit der Verschriftungsproblematik verbunden ist die Typographieproblematik der Vorlagen; die verwendeten Typographien spiegeln die Drucktechnik zu Ende des vorvergangenen Jahrhunderts und deren Entwicklung über einen Publikationszeitraum von mehr als 20 Jahren wider. Genau diese Probleme sind aber nicht allein mit dem Bündnerromanischen verbunden, sondern finden sich ähnlich in anderen Sprachen, auch in Kleinsprachen. Insofern ist die Lösung der anstehenden Probleme exemplarisch für die Tiefenerschließung von Textkorpora und die Bewahrung des kulturellen Erbes von Sprachgemeinschaften. Das DRC-Projekt betritt dabei methodisches Neuland: Es bindet die Sprachgemeinschaft über das Internet in die Erschließung ein. Dies betrifft in gleicher Weise die Erstellung wie auch die Nutzung der gewonnenen Volltexte. Die Sprachge-

[10] Cf. <www.digizeitschriften.de/>.

[11] Zu korpuslinguistischen Potentialen cf. u.a. Mcenery/Wilson 2001.

meinschaft partizipiert an der Korrektur, der Pflege und der anreichernden Auszeichnung und an der Erweiterung der Textbasis. Dies verleiht dem Projekt einen prototypischen Charakter für die Erstellung digitaler Korpora auf Grundlage gedruckter Textsammlungen, gerade für kleinere Sprachgemeinschaften mit fehlenden oder varianten Verschriftungstraditionen (z.B. Aromunisch oder Sardisch).

Um die Digitalisate – d.h. die Bildvorlagen – für eine textuelle Verarbeitung zugänglich zu machen, müssen sie in eine maschinenlesbare Form überführt werden. Hierfür werden die erstellten Bildvorlagen durch optische Zeichenerkennung (*Optical Character Recognition*, OCR) bearbeitet und in Text umgewandelt. Im Anschluss an die OCR-Erfassung werden die Texte einem mehrschrittigen Korrekturvorgang unterzogen. Die Notationen folgen einem XML-Standard. Dies ermöglicht auch zukünftig den Zugriff und die Weiterverarbeitung der Texte durch offene und frei zugängliche Werkzeuge. Die annotierten Texte werden in einer öffentlich zugänglichen Datenbank abgelegt. Korrekturen werden im Rahmen eines auch in der Versionierung an Wiki-Technologien angelehnten Verfahrens interaktiv und kollaborativ unter Einbeziehung von Angehörigen und Interessierten der Sprachgemeinschaft durchgeführt. Dadurch werden die bekannten Probleme der OCR bei älteren, nicht normierten und typographisch varianten Schriftsystemen abgefangen; über das konkrete materielle Ziel hinaus werden damit übertragbare und somit nachhaltige, kompetenzorientierte Verfahren entwickelt, die für die Tiefendigitalisierung des schriftlichen kulturellen Erbes kleinerer Sprachgemeinschaften typisch sind. Von besonderem Interesse ist die Möglichkeit für Mitglieder solcher Sprachgemeinschaften, über Wiki-ähnliche Technologien den Erhalt des sprachlichen und kulturellen Erbes aktiv zu unterstützen. Dass solche Technologien großen Anklang in Sprechergemeinschaften finden, zeigt u.a. die bündnerromanische Wikipedia,[12] die ca. 3.300 Artikel umfasst und derzeit mehr als 5.200 Nutzer ausweist (Stand jeweils 02.03.2012). Damit partizipieren ca. 10% der Sprachgemeinschaft an dem Wikipediaprojekt.

Teil des Vorhabens ist auch die Entwicklung spezialisierter Korrekturverfahren in vernetzten Systemen. Hierfür steht der Sprachlichen Informationsverarbeitung mit "Tesla"[13] eine geeignete informationstechnologische Infrastruktur zur Verfügung. "Tesla" ist ein Komponentensystem zur Prozessierung von Texten. Die graphisch konfigurierbaren Komponenten ermöglichen auch Nichtpro-

[12] Cf. <rm.wikipedia.org/wiki/Pagina_principala>.

[13] *Text Engineering Software Laboratory*, cf. <www.spinfo.uni-koeln.de/forschung/tesla.html>.

grammierern, Komponenten zu Prozessketten zu konfigurieren und Texte zu verarbeiten. Dahinter steht die Umsetzung der Idee des Programmierens im Großen mit Hilfe vorhandener Komponenten, die wie Lego-Steine zusammengesetzt werden.

Im Zusammenhang mit der Chrestomathie kann zum einen auf die in "Tesla" vorhandenen sprachverarbeitenden Komponenten zurückgegriffen werden, zum anderen werden im Rahmen von Begleitforschungen neue Komponenten erstellt und eingebunden.

Neben Korrekturverfahren werden Methoden zur korpuslinguistischen Aufbereitung eingesetzt und weiterentwickelt. Ziel ist es, die Volltexte einerseits als annotiertes Korpus, andererseits als Textsammlung für strukturierten Zugriff (u.a. Volltextsuche) zur Verfügung zu stellen. Die für entsprechende Retrieval-Verfahren grundlegenden Techniken wie Indexierung, Konkordanzen etc. werden von der Sprachlichen Informationsverarbeitung bereitgestellt und im Rahmen des Projekts auf die speziellen Anforderungen der vorliegenden Textsammlung hin weiterentwickelt. Der Übergang von Texterstellung, -anreicherung und -nutzung – d.h. zur Nachnutzung – ist dabei fließend. Dies liegt in der Natur digitaler Daten, die im Gegensatz zu Printdaten sowohl lesbar als auch veränderbar bzw. schreibbar sind. Die Algorithmen zur Textaufbereitung und zur Textkorrekur sind auch für die Nutzung einzusetzen; dies gilt z.B. für Indizierungs-, Konkordanz-, Such- oder Musterbildungsalgorithmen. Die Algorithmen stehen unter dem "Tesla"-System als Komponenten zur Verfügung.

Die im Vorhaben eingesetzten Erschließungs- und Auszeichnungstechniken sollen in der Folge auf weitere Textsammlungen des Bündnerromanischen angewendet werden, um den Bestand an Korpora zu erhöhen. Als Quelle für weitere bündnerromanische Texte bieten sich etwa die alten Bibeldrucke an.

Alle Korpora werden zum einen einzeln durch die beteiligten Institutionen, zum anderen als Bestandteil des Komponentensystems "Tesla" für korpuslinguistische Untersuchungen der Forschergemeinschaft frei zur Verfügung gestellt. Die Einbindung darin bringt mit sich, dass zu den Korpora auch die eingesetzten Auszeichnungtechniken (Algorithmen, Verfahren) mitgeliefert werden können. Dadurch profitieren nicht nur Linguisten: Auch Literatur- und Kulturwissenschaftlern stehen neue Recherche- und Arbeitsmöglichkeiten zur Verfügung. Spezielle korpuslinguistische, insbesondere sprachübergreifende und sprachvergleichende Untersuchungen werden damit langfristig für das Bündnerromanische möglich.

Die Digitalisate wurden von dem Digizeitschriften-Projekt der SUB (Staats- und Universitätsbibliothek) Göttingen freundlicherweise mit den zugehörigen Metadaten bereitgestellt. Die Volltexte des Chrestomathieprojekts werden an das Digizeitschriften-Projekt zurückfließen und die SUB Göttingen darin unterstützen, digitalisierte Zeitschriften für rasche Informationsgewinnung zu erschließen.

5. Ähnliche Projekte

Aus der großen Zahl von Digitalisierungs- und Tiefenerschließungsprojekten werden hier einige ausgewählte vorgestellt. Auswahlkriterien sind die jeweiligen Gemeinsamkeiten und spezifischen Differenzen, die abschließend synoptisch dargestellt werden. Bei den Gemeinsamkeiten gilt das Augenmerk vor allem der OCR-Problematik und ihren Lösungsansätzen.

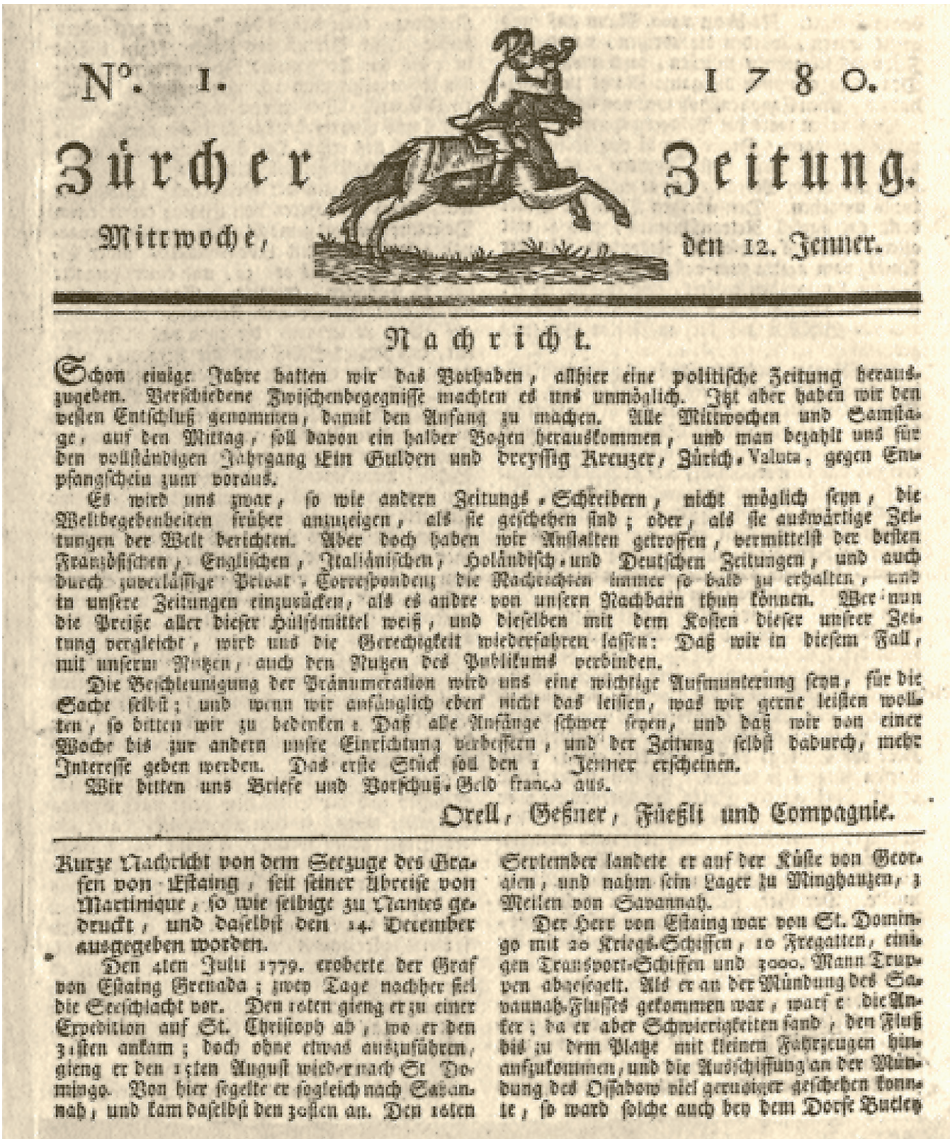
5.1 Die "Neue Zürcher Zeitung"

Die "Neue Zürcher Zeitung" (NZZ) ist eine der ältesten Zeitungen der Welt. Sie erscheint seit 1780 (bis 1821 als "Zürcher Zeitung"). Zu ihrem 225-jährigen Bestehen – 2005 – initiierte sie das Projekt "Archiv 1780", um das bis dahin mikroverfilmte Archiv zu digitalisieren, OCR-zu lesen und damit der elektronischen Volltextrecherche zugänglich zu machen. Der Umfang des Vorhabens ist sehr beeindruckend, handelt es sich doch um mehr als 1.500 Mikrofilme, die über zwei Millionen Zeitungsseiten darstellen.

Die NZZ arbeitete bei diesem Projekt mit dem Institut für Medienkommunikation der Fraunhofer-Gesellschaft in St. Augustin (bei Bonn) zusammen. In einem ersten Schritt wurden die Digitalisate weitestgehend automatisch aufgearbeitet. Der nächste und schwierige Schritt war die OCR-Erfassung. Hier kam die Typographieproblematik ins Spiel. Bis 1946 verwendete die NZZ eine Frakturschrift, ab dann setzte sie Antiqua ein.

Für die OCR-Erfassung kam der ABBYY-*FineReader* zum Einsatz, speziell auch im Hinblick auf die Fähigkeit zum Lesen von Frakturschriften. Angesichts der immensen Datenmengen setzte das Projekt auf eine möglichst hohe Trefferquote bei der OCR-Lektüre und verzichtete auf manuelle Nachbereitung.

Der Zugriff auf das Archiv erfolgt über ein Web-Interface, das mit einer leistungsfähigen Datenbank verknüpft ist. Die Texte sind lesbar, aber nicht veränderbar *(read-only)*. Die Texte werden durch Metadaten ausgezeichnet, auf die für die Recherche zugegriffen werden kann.



N°. 1. 1780.

Zürcher Zeitung.

Mittwoche, den 12. Jenner.

Nachricht.

Schon einige Jahre hatten wir das Vorhaben, allhier eine politische Zeitung herauszugeben. Verschiedene Zwischenbegegnisse machten es uns unmöglich. Itzt aber haben wir den vesten Entschluß genommen, damit den Anfang zu machen. Alle Mittwochen und Samstage, auf den Mittag, soll davon ein halber Bogen herauskommen, und man bezahlt uns für den vollständigen Jahrgang Ein Gulden und dreyssig Kreuzer, Zürich-Valuta, gegen Empfangschein zum voraus.

Es wird uns zwar, so wie andern Zeitungs-Schreibern, nicht möglich seyn, die Weltbegebenheiten früher anzuzeigen, als sie geschehen sind; oder, als sie auswärtige Zeitungen der Welt berichten. Aber doch haben wir Anstalten getroffen, vermittelst der besten Französischen, Englischen, Italiänischen, Holändisch- und Deutschen Zeitungen, und auch durch zuverläßige Privat-Correspondenz die Nachrichten immer so bald zu erhalten, und in unsere Zeitungen einzurücken, als es andre von unsern Nachbarn thun können. Wer nun die Preiße aller dieser Hülfsmittel weiß, und dieselben mit dem Kosten dieser unserer Zeitung vergleicht, wird uns die Gerechtigkeit wiederfahren lassen: Daß wir in diesem Fall, mit unserm Nutzen, auch den Nutzen des Publikums verbinden.

Die Beschleunigung der Pränumeration wird uns eine wichtige Aufmunterung seyn, für die Sache selbst; und wenn wir anfänglich eben nicht das leisten, was wir gerne leisten wollten, so bitten wir zu bedenken: Daß alle Anfänge schwer seyen, und daß wir von einer Woche bis zur andern unsre Einrichtung verbessern, und der Zeitung selbst dadurch, mehr Interesse geben werden. Das erste Stück soll den 1 Jenner erscheinen.

Wir bitten uns Briefe und Vorschuß-Geld franco aus.

Orell, Geßner, Füeßli und Compagnie.

Kurze Nachricht von dem Seezuge des Grafen von Estaing, seit seiner Abreise von Martinique, so wie selbige zu Nantes gedruckt, und daselbst den 14. December ausgegeben worden.

Den 4ten Julii 1779. eroberte der Graf von Estaing Grenada; zwey Tage nachher fiel die Seeschlacht vor. Den 10ten gieng er zu einer Expedition auf St. Christoph ab, wo er den 31sten ankam; doch ohne etwas auszuführen, gieng er den 15ten August wieder nach St. Domingo. Von hier segelte er sogleich nach Savannah, und kam daselbst den 30sten an. Den 16ten September landete er auf der Küste von Georgien, und nahm sein Lager zu Minghauzen, 3 Meilen von Savannah.

Der Herr von Estaing war von St. Domingo mit 20 Kriegs-Schiffen, 10 Fregatten, einigen Transport-Schiffen und 3000. Mann Truppen abgesegelt. Als er an der Mündung des Savannah-Flusses gekommen war, warf er die Anker; da er aber Schwierigkeiten fand, den Fluß bis zu dem Platze mit kleinen Fahrzeugen hinaufzukommen, und die Ausschiffung an der Mündung des Offadow viel geruhiger geschehen konnte, so ward solche auch bey dem Dorfe Bucley

Fig. 2: Digitalisat der NZZ
(aus: <www.iais.fraunhofer.de/uploads/media/NZZ_fhg_journal_imk_nzz.pdf>)

Die Charakteristika des NZZ-Archiv-Projekts sind (im Hinblick auf das Chrestomathie-Projekt) die Datenmengen, die Typographieproblematik, das Webinterface und die Datenbanktechnologie.[14]

[14] Cf. <www.iais.fraunhofer.de/uploads/media/NZZ_fhg_journal_imk_nzz.pdf>, <www.iais.fraunhofer.de/uploads/media/NZZ_225_jahrgaenge_originalartikel.pdf>

5.2. "Australian Newspapers Digitisation Program"

Das "Australian Newspapers Digitisation Program" (ANDP) (<www.nla.gov.au/ndp/>) ist ein von der "National Library of Australia" durchgeführtes Projekt zur Digitalisierung und Volltexterschließung aller australischen Zeitungen. Bis Mitte 2011 stehen ca. vier Millionen Zeitungsseiten zur Verfügung, die einen Zeitraum von 1803 bis 1954 umfassen. Im September 2008 standen bereits ca. 2,5 Millionen Artikel auf 250.000 digitalisierten Zeitungsseiten *online* zur Verfügung.

Das spezifische Merkmal dieses Projekts ist neben der beeindruckenden Datenmenge der Community-orientierte Ansatz der Fehlerkorrektur der OCR-gelesenen Texte. Die Leser im Web haben Schreibzugriff auf die Texte, um Fehler der OCR korrigieren zu können. Das ANDP-Projekt wurde mit einer besonderen Typographieproblematik konfrontiert, die die OCR erschwerte. Für den Druck australischer Zeitungen wurden oftmals Satzmaschinen verwendet, die im Mutterland England ausgemustert worden waren. Dies verminderte die Qualität der Typographie und erhöhte die Fehlerrate bei der OCR. Um Fehler korrigieren zu können, setzte das ANDP auf die vernetzte Webgemeinde und gab ihr für die Korrektur Schreibzugriff. Das Hauptinteresse der Webcommunity ist übrigens stark genealogisch motiviert vor dem Hintergrund der

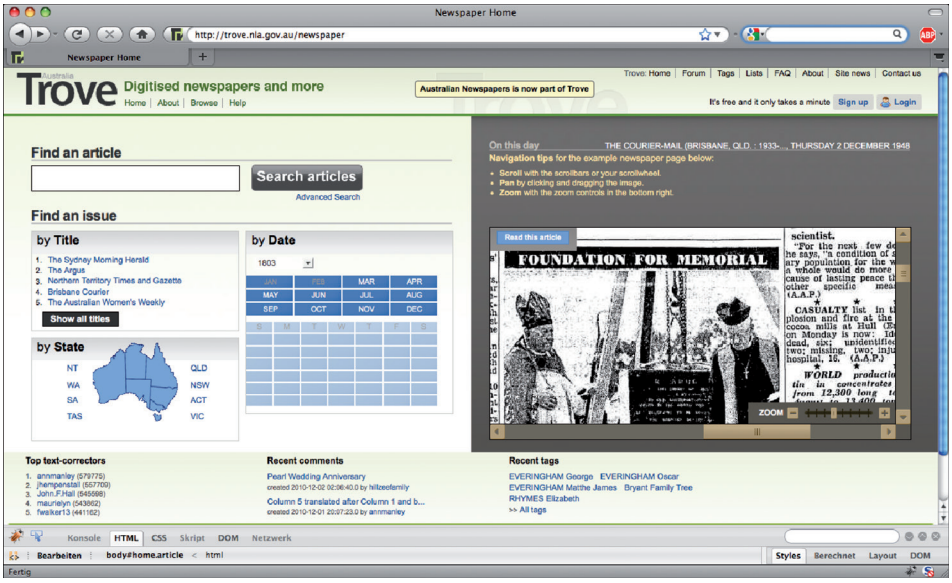


Fig. 3: ANDP (cf. <www.trove.nla.gov.au/newspaper>)

Frage, wann und gegebenenfalls unter welchen Bedingungen die Vorfahren in Australien eingewandert sind:[15]

> Through this service the Library is providing access to every article, advertisement and illustration on every newspaper page being digitised. Users can browse the newspaper pages or search across the full text of the articles. Key users of the service to date, range from academics, family historians and social and economic researchers through to school students.[16]

5.3. "Text+Berg"

Der 1863 gegründete Schweizer Alpen-Club (SAC) publiziert seine Jahrbücher seit 1864. Das "Jahrbuch des S.A.C." erschien von 1864 bis 1923, sein französischsprachiges Pendant unter dem Titel "Echo des Alpes" von 1872 bis 1923. Von 1925 bis heute erscheint das Jahrbuch des SAC unter dem Titel "Die Alpen".

Das von Martin Volk (Universität Zürich) geleitete Projekt "Text+Berg" (<www.textberg.ch>) erschließt seit 2008 die Inhalte der Zeitschriften. Deren Texte werden nach der Digitalisierung OCR-gelesen, automatisch korrigiert und korpuslinguistisch aufbereitet. Die korpuslinguistische Aufbereitung identifiziert die Sprache der Texte (Deutsch, Französisch, Italienisch, auch Rätoromanisch). Die Identifikation der Sprache ist insofern nicht trivial, als des Öfteren gemischtsprachige Texte vorliegen, z.B. ein deutschsprachiger Text, in dem sich wörtliche Zitate eines französischsprachigen Bergführers finden. Für den Zugriff werden den Texten Metainformationen hinzugefügt. Von besonderem Wert sind die Verfahren zur Erkennung von Toponymen, die geographische Bezeichnungen der Texte mit der Datenbank der Schweizer Landeskarten verknüpfen.

Die Schriften des SAC sind in ihrer 1864 ansetzenden Geschichte eine einzigartige, mehrsprachige Textbasis von großem historischen, kulturwissenschaftlichen und sprachwissenschaftlichen Wert.

Inhaltlich spiegeln sie für das Chrestomathieprojekt eine geographische und mitunter auch inhaltliche Nähe wider. Methodisch sind die zum Einsatz gekommenen computerlinguistischen Verfahren – sowohl zur Erstellung der

[15] Cf. z.B. <libraryroad.wordpress.com/2010/08/01/the-australian-newspaper-digitisation-program-a-model-of-user-library-collaboration/>, <www.nla.gov.au/anplan>.

[16] <http://www.nla.gov.au/ndp/news_and_events/documents/ANDPSuccessstory_OnlineCurrents-Dec2009.pdf>, 1.

Textbasis als auch für den Zugriff darauf – für das DRC-Projekt von großem Interesse. Da die Träger des Züricher "Text+Berg" Projekts ihrerseits an den für das DRC-Projekt entwickelten communitybasierten Korrektur- und Auszeichnungsverfahren interessiert sind – vor dem Hintergrund von ca. 120.000 Mitgliedern des SAC, von denen ein Teil als Korrektoren gewonnen werden könnten – vereinbarten die *Sprachliche Informationsverarbeitung* und das *Zürcher Institut für Computerlinguistik* Anfang 2011 Kooperation und Austausch.

5.4. Korpora zum Dolomitenladinischen: TALL

Die Erstellung von Korpora zum Dolomitenladinischen (TALL: *Tratament Automatich dl Lingaz Ladin*) weist sprachlich und thematisch die größte Nähe zum DRC-Projekt auf. Dies ist sicher vom Gegenstand her begründet: Von jeher fasste eine integralistische Romanistik das Bündnerromanische, das Dolomitenladinische und das Friulanische sprachlich und im Sprachbewusstsein durchaus als Einheit (hier ist nicht der Ort, die *questione ladina* zu repetieren). Quantitative Untersuchungen haben die Fragestellungen methodisch bereichert und datengetrieben gesichert. Vor einem solchen Hintergrund stellt das TALL-Projekt mit

Fig. 4: Das dolomitenladinische TALL-Projekt: <http://vll.ladintal.it/applications/textanalysis/site-bolzano/index_de.jsp> (letzte Abfrage: 20.9.2012)

seinen Bezügen zu einer varietäten- und kodifikationsreichen Minderheitensprache eine sozusagen ältere Schwester zum DRC-Projekt dar.

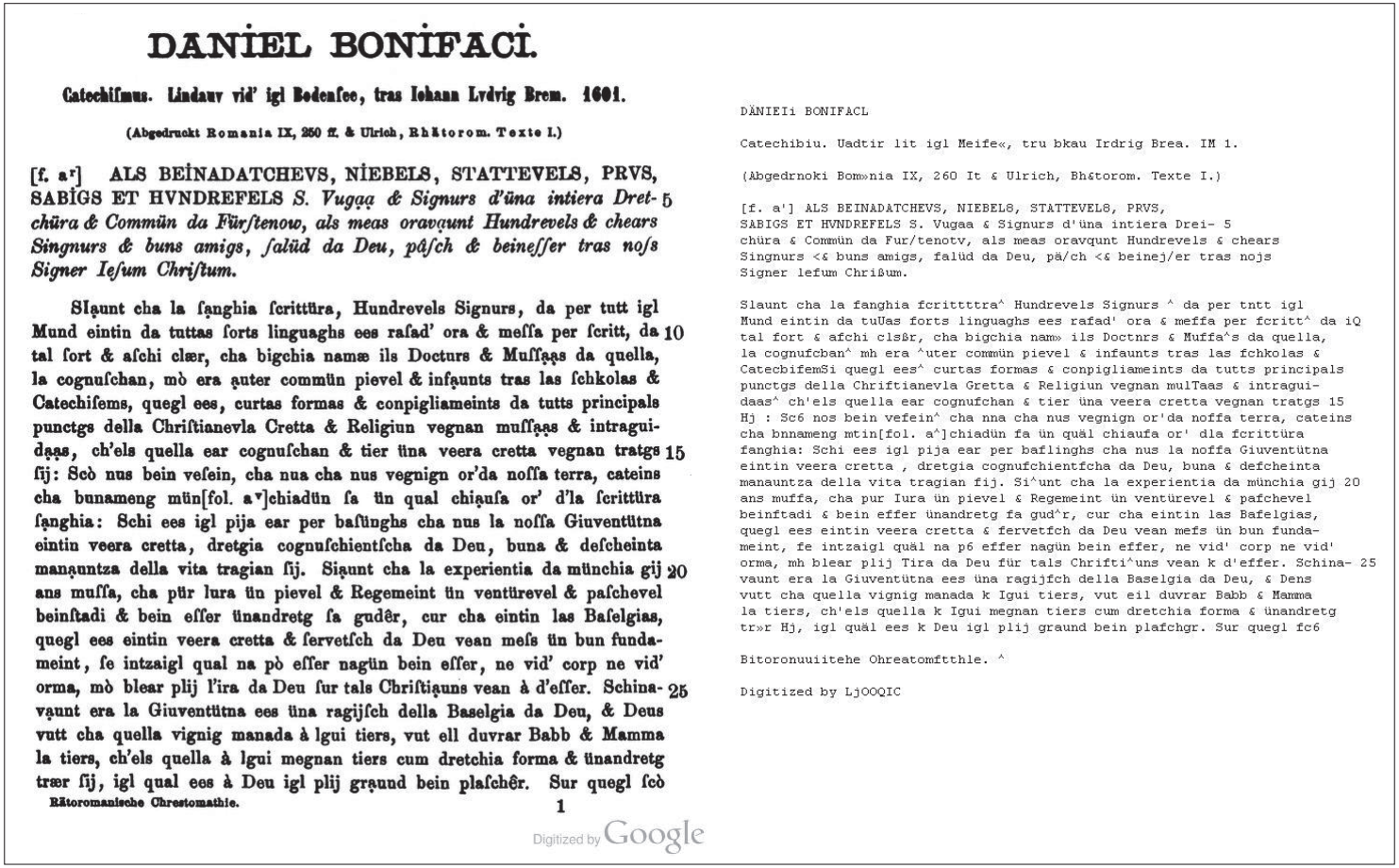
Damit ergibt sich auch die Möglichkeit, aus Vergleich und Auswertung der Korpora die Untersuchungen zum Rätoromanischen datengetrieben und empirisch (d.h. über das Schicksal von z.B. lat. AU- , dem "Arbeitspferd" historisch-phonetischer Betrachtung, hinaus) in den verschiedenen linguistischen Disziplinen (Lexikologie, Morphologie, Syntax, Semantik) zu sichern und durch neue Fragestellungen zu erweitern.

Im methodischen Vorgehen, auch bedingt durch die zugrundeliegenden Quellen, sind die beiden Projekte unterschiedlich angelegt. Daraus ergibt sich die Gelegenheit, beide Projekte zusammenzuführen, um wechselseitig zu profitieren. Dies betrifft nicht nur die Daten der Korpora, sondern auch die Methoden und deren Implementationen.

5.5. *Google*

Das Projekt *Google Books* ist ohne Zweifel das größte, geradezu das gigantischste Digitalisierungs- und OCR-Projekt. Schätzungen gehen davon aus, dass bislang ca. 15 Prozent des weltweiten Buchbestands durch *Google* erfasst worden sind. Zugriff und Verarbeitungsmöglichkeiten verändern sich durch *Google* in revolutionärer Weise. Das Vorgehen von *Google* ist stark pragmatisch und mitunter für Philologen befremdlich. Korrekturen von OCR-gelesenen Texten werden nicht vorgenommen. *Google Books* hat auch – bei der Digitalisierung der romanistischen Bestände der Universitätsbibliothek von Harvard – die Zeitschrift "Romanische Forschungen" und somit die Rätoromanische Chrestomathie digitalisiert und OCR-gelesen. Dabei wurde folgendes Ergebnis erzielt:

Der Text von *Google* ist aufgrund seiner sehr hohen Fehlerzahl unbrauchbar. Hier muss jedoch darauf verwiesen werden, dass der vorgestellte Text der Rätoromanischen Chrestomathie für OCR-Erfassung sehr schwierig ist. Dies liegt vor allem an der Typographie und der Sprache. Die Ergebnisse von *Google* beispielsweise für modernere surselvische Texte sind weniger katastrophal, gleichwohl aber genauso wenig brauchbar.



DANIEL BONIFACI.

Catechiſmus. Lindauv vid' igl Bodenſee, tras Iohann Lvdvig Brem. 1601.

(Abgedruckt Romania IX, 250 ff. & Ulrich, Rhätorom. Texte I.)

[f. a^{r}] ALS BEINADATCHEVS, NIEBELS, STATTEVELS, PRVS, SABIGS ET HVNDREFELS *S. Vugaꝙ & Signurs d'üna intiera Dretchüra & Commün da Fürſtenow, als meas oravꝙunt Hundrevels & chears Singnurs & buns amigs, ſalüd da Deu, pâſch & beineſſer tras noſs Signer Ieſum Chriſtum.*

Slaunt cha la ſanghia ſcritttüra, Hundrevels Signurs, da per tutt igl Mund eintin da tuttas ſorts linguaghs ees raſad' ora & meſſa per ſcritt, da tal ſort & aſchi clær, cha bigchia namæ ils Docturs & Muſſaas da quella, la cognuſchan, mò era auter commün pievel & infaunts tras las ſchkolas & Catechiſems, quegl ees, curtas formas & conpigliameints da tutts principals punctgs della Chriſtianevla Cretta & Religiun vegnan muſſaas & intraguidaas, ch'els quella ear cognuſchan & tier üna veera cretta vegnan tratgs ſij: Scò nus bein veſein, cha nua cha nus vegnign or'da noſſa terra, cateins cha bunameng mün[fol. a^{v}]chiadün ſa ün qual chiauſa or' d'la ſcrittüra ſanghia: Schi ees igl pija ear per baſünghs cha nus la noſſa Giuventütna eintin veera cretta, dretgia cognuſchientſcha da Deu, buna & deſcheinta manauntza della vita tragian ſij. Siaunt cha la experientia da münchia gij ans muſſa, cha pür lura ün pievel & Regemeint ün ventürevel & paſchevel beinſtadi & bein eſſer ünandretg ſa gudêr, cur cha eintin las Baſelgias, quegl ees eintin veera cretta & ſervetſch da Deu vean meſs ün bun fundameint, ſe intzaigl qual na pò eſſer nagün bein eſſer, ne vid' corp ne vid' orma, mò blear plij l'ira da Deu ſur tals Chriſtiauns vean à d'eſſer. Schinavaunt era la Giuventütna ees üna ragijſch della Baſelgia da Deu, & Deus vutt cha quella vignig manada à lgui tiers, vut ell duvrar Babb & Mamma la tiers, ch'els quella à lgui megnan tiers cum dretchia forma & ünandretg trær ſij, igl qual ees à Deu igl plij graund bein plaſchêr. Sur quegl ſcò

DÀNIEIi BONIFACL

Catechibiu. Uadtir lit igl Meife«, tru bkau Irdrig Brea. IM 1.

(Abgedrnoki Bom»nia IX, 260 It & Ulrich, Bh&torom. Texte I.)

[f. a'] ALS BEINADATCHEVS, NIEBEL8, STATTEVEL8, PRVS,
SABIGS ET HVNDREFELS S. Vugaa & Signurs d'üna intiera Drei- 5
chüra & Commün da Fur/tenotv, als meas oravqunt Hundrevels & chears
Singnurs <& buns amigs, falüd da Deu, pä/ch <& beinej/er tras nojs
Signer lefum Chrißum.

Slaunt cha la fanghia fcritttttra^ Hundrevels Signurs ^ da per tntt igl
Mund eintin da tuUas forts linguaghs ees rafad' ora & meffa per fcritt^ da iQ
tal fort & afchi clsßr, cha bigchia nam» ils Doctnrs & Muffa^s da quella,
la cognufcban^ mh era ^uter commün pievel & infaunts tras las fchkolas &
CatecbifemSi quegl ees^ curtas formas & conpigliameints da tutts principals
punctgs della Chriftianevla Gretta & Religiun vegnan mulTaas & intraguidaas^ ch'els quella ear cognufchan & tier üna veera cretta vegnan tratgs 15
Hj : Sc6 nos bein vefein^ cha nna cha nus vegnign or'da noffa terra, cateins
cha bnnameng mtin[fol. a^]chiadün fa ün quäl chiaufa or' dla fcrittüra
fanghia: Schi ees igl pija ear per baflinghs cha nus la noffa Giuventütna
eintin veera cretta , dretgia cognufchientfcha da Deu, buna & defcheinta
manauntza della vita tragian fij. Si^unt cha la experientia da münchia gij 20
ans muffa, cha pur Iura ün pievel & Regemeint ün ventürevel & pafchevel
beinftadi & bein effer ünandretg fa gud^r, cur cha eintin las Bafelgias,
quegl ees eintin veera cretta & fervetfch da Deu vean mefs ün bun fundameint, fe intzaigl quäl na p6 effer nagün bein effer, ne vid' corp ne vid'
orma, mh blear plij Tira da Deu für tals Chrifti^uns vean k d'effer. Schina- 25
vaunt era la Giuventütna ees üna ragijfch della Baselgia da Deu, & Dens
vutt cha quella vignig manada k Igui tiers, vut eil duvrar Babb & Mamma
la tiers, ch'els quella k Igui megnan tiers cum dretchia forma & ünandretg
tr»r Hj, igl quäl ees k Deu igl plij graund bein plafchgr. Sur quegl fc6

Bitoronuuiitehe Ohreatomftthle. ^

Fig. 5: Digitalisat und OCR-erfasster Text aus *Google Books* (cf. <http://www.archive.org/details/rtoromanischech00camigoog>)

5.6. Spezifische Merkmale und spezifische Differenzen

Abschließend werden die ausgewählten Digitalisierungsprojekte in einer synoptischen Sicht gegenübergestellt. Die Übersicht stellt die spezifischen Merkmale und Differenzen heraus. Allen Projekten ist gemeinsam, dass die Texte digitalisiert und OCR-gelesen werden, dass also auf sogenannte *Double-Keying*-Verfahren – ein Euphemismus für buchstabenweises, zweifaches Abtippen in Niedriglohnländern – verzichtet wird.

Ein "+" in der Zeile zu Sprachen wird sowohl beim Bündnerromanischen und beim Dolomitenladinischen gesetzt: Darin sind Varietäten mit unterschiedlichen Verschriftungstraditionen und -konventionen eingeschlossen.

Schließlich soll die Übersicht auch zeigen, dass das DRC-Projekt durch die gemeinschaftsorientierte Korrektur und Anreicherung neue Wege geht. Dies zeigen die Zeilen mit den Merkmalen Korrektur und Anreicherung. Hier sind Korrektur und Anreicherung durch die Sprachgemeinschaft und durch an der Sprache Interessierte gemeint. Das Projekt setzt auf webbasierte Partizipation. Dies scheint uns ein Novum für Minderheitensprachen zu sein. Traditionelle (Massen-)Medien sind für Kleinsprachen oft eine Bedrohung; deren Sprache ist nicht die Minderheitensprache, sondern die der Mehrheit, der großen Sprechergruppen (z.B. in der Schweiz) oder der Nationalsprachen. Massenmedien bedürfen – es klingt trivial – der Masse, um Einkünfte durch Leser-, Hörer- oder Zuschauerzahl auch mit Hilfe der Werbung zu sichern. Kleine, oftmals fragmentierte Sprachgemeinschaften stellen keine Masse dar. Minderheitengemeinschaften sind fast von Natur aus individualisiert. Die neuen, sozialen Medien sind viel geeignetere Instrumente: sie individualisieren Kommunikation und ermöglichen die aktive, gestaltende Teilhabe.

Im Hinblick auf die Rätoromanische Chrestomathie geht es nicht allein um die Korrektur und darüber hinaus die Anreicherung eines Textes im Web, sondern zusätzlich um die Kommunikation der Mitglieder der Sprachgemeinschaft(en) untereinander. Diese Kommunikation hat die Objektsprache (der Texte) zum Inhalt, sie ist also auch eine aktive Auseinandersetzung und Reflexion der (Mutter-) Sprache und idealerweise eine Erweiterung des Sprachbewusstseins.

Dies sei in der nachfolgenden Übersicht unter Stichpunkt "Anreicherung" zu berücksichtigen; Anreicherung des Textes setzt Auseinandersetzung und Kommunikation voraus.

	Google	NZZ	ANDP	Text+Berg	TALL	DRC
OCR	+	+	+	+	+	+
Quantität	++	+	+	+	–	–
Sprachen	++	-	-	+	+	+
Korrektur	-	-	+	+-	+-	++
Anreicherung	-	-	-	-	-	+

Fig. 6: Gegenüberstellung der Digitalisierungsprojekte

6. Methodisches Vorgehen

6.1 Übersicht

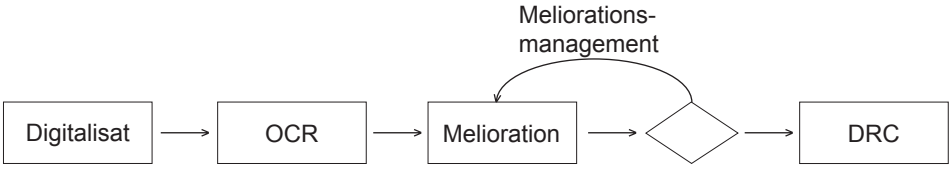


Fig. 7: *Workflow* DRC

Die Skizze zeigt den Workflow bei der Erstellung der DRC. Die Digitalisate kommen zum größeren Teil aus dem Digizeitschriftenprojekt (s.u.), zum kleineren Teil werden sie selbst gescannt. Die Digitalisate werden dann OCR-gelesen. Im darauf folgenden iterativen Verfahren werden die OCR-Texte korrigiert, annotiert und angereichert. Dieser Vorgang – der über bloße Korrektur hinausgeht – wird "Melioration" genannt. Das geschieht unter der Anleitung eines Moderators (Florentin Lutz) webbasiert durch die Sprachgemeinschaft und durch am Bündnerromanischen Interessierte. Die Texte der Chrestomathie stehen danach als Volltexte im Web zur Verfügung (<www.crestomazia.ch>).

6.2 Digitalisate als Ausgangsdaten

Die Rätoromanische Chrestomathie erschien zwischen 1896 und 1919 in den "Romanischen Forschungen" (s.o.). Die "Romanischen Forschungen" wurden im Rahmen des von der Deutschen Forschungsgemeinschaft geförderten Projekts Digizeitschriften von der Staats- und Universitätsbibliothek Göttingen digitalisiert. Auf die Digitalisate kann unter <www.digizeitschriften.de> über das Web zugegriffen werden:

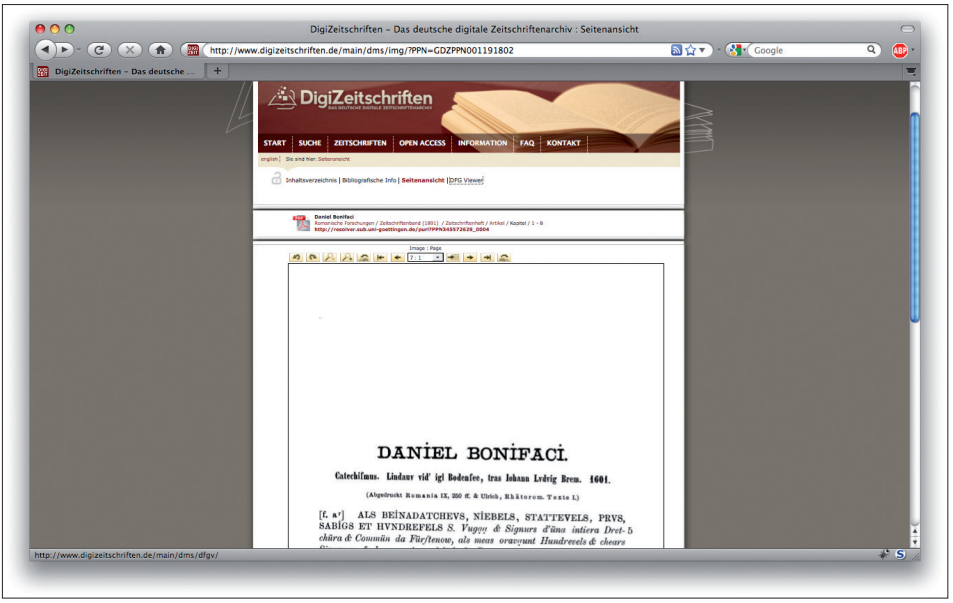


Fig. 8: Das Projekt Digizeitschriften der SUB Göttingen

Die Staats- und Universitätsbibliothek Göttingen stellte die Digitalisate dem DRC-Projekt als hochauflösende Tiffs zur Verfügung. Zusätzlich stellte die SUB Göttingen auch die Metadaten zur Verfügung, die eine der Grundlagen für die Anreicherung und Auszeichnung der Texte bilden.

Weitere Digitalisate – z.B. die Bände XIV und XV der Octopus-Ausgabe von 1985 und 1986 – werden mit Hilfe des Scanroboters der Universitäts- und Stadtbibliothek Köln erzeugt. Dabei handelt es sich um einen Hochleistungsscanner, der automatisch das Umschlagen der Seiten besorgt. Der Roboter verfügt über vielfältige Einstellungsmöglichkeiten, um beispielsweise unterschiedliche Buchgrößen oder

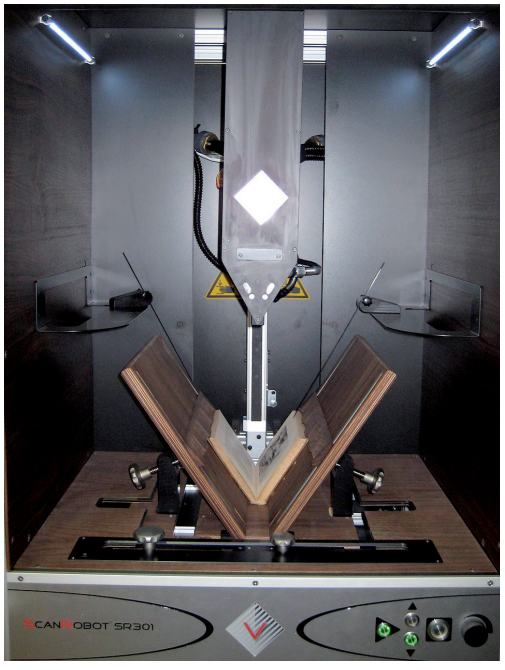


Fig. 9: Scanroboter der Universitäts- und Stadtbibliothek Köln

Papierarten schonend bei der Digitalisierung zu bearbeiten. Die zu digitalisierenden Bücher werden in eine Buchwippe gelegt.

6.3. Optische Zeichenerkennung

Die Digitalisate bilden die Eingabe für die optische Zeichenerkennung (OCR: *Optical Character Recognition*). Die Zeichenerkennung setzt zwei Bearbeitungsschritte voraus: Zum einen die Erkennung bzw. manuelle Markierung von Textabschnitten, zum anderen das Erlernen und das Trainieren von Zeichenmustern.

Die Frage der Markierung von Textabschnitten stellt sich vor allem bei mehrspaltigen Seiten, wie das folgende Beispiel (cf. Fig. 10), in dem die (grünen) Rahmen durch manuellen Eingriff gesetzt wurden, zeigt.

Für die Verminderung der Lesefehler ist das sorgfältige Trainieren von Zeichenmustern sehr wichtig. Die Rätoromanische Chrestomathie – über einen Zeitraum von mehr als 20 Jahren veröffentlicht – ist von hoher typographischer Vielfalt. Dies ist eine Herausforderung für die OCR, um so mehr, als der Zeichenerkennung keine angemessenen Korrekturlexika für die verschiedenen Idiome zur Verfügung stehen. Gerade die älteren Texte der Chrestomathie sind orthographisch (noch) nicht normiert. Die Idiome des Bündnerromanischen folgen unterschiedlichen Verschriftungsformen und -traditionen. Daher kommt dem Mustertraining hohe Bedeutung zu.

Für die OCR hat sich der ABBYY-*FineReader* als am besten geeignet erwiesen. Als Ausgabe erzeugt der *Finereader* PDF-Dateien, in denen die Bilder (i.e. die Digitalisate) der Textvorlage und die erkannten Texte eingebettet sind. Die PDF-Dateien enthalten Angaben über die Positionen der erkannten Zeichenketten in den Bilddateien. Dadurch kann berechnet werden, wo die erkannte Zeichenkette in der Bilddatei vorkommt. Dies ist für die Unterstützung der im nächsten Abschnitt beschriebenen Melioration von hoher Wichtigkeit, da das Meliorationswerkzeug Bilder, Text und anreichernde Annotationen zusammenführt.

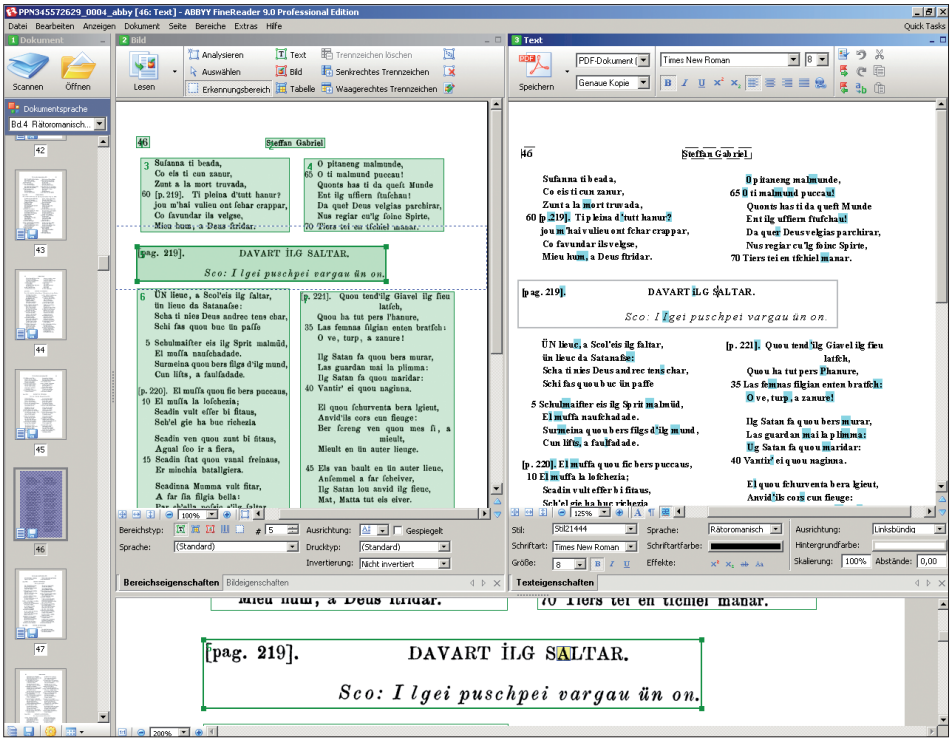


Fig. 10: Spaltenmarkierung und OCR

6.4. Melioration

Die Melioration ist ein iterativer, interaktiver Prozess, der die Korrektur und die anreichernde Auszeichnung des OCR-erstellten Textes und die Interaktion der Melioratoren untereinander und mit dem Moderator umfasst. Die Melioration erfordert ein webbasiertes Editorwerkzeug. Dieses Werkzeug ermöglicht es, die OCR-gelesenen Texte zu korrigieren und mit Metainformationen auszuzeichnen. Eine wesentliche Funktion des Editorwerkzeugs ist es, nicht nur den zu korrigierenden Text, sondern auch das Inputimage zu präsentieren, aus dem mit Hilfe des OCR-Programms der Text erzeugt wurde. Die Arbeit der Melioratoren wird zusätzlich dadurch unterstützt, dass Text und Bild direkt verknüpft sind: Wird im OCR-gelesenen Text eine möglicherweise unsichere Form gefunden und mit der Maus angeklickt, so wird im Image das zugrundeliegende Buchstabenbild durch einen kleinen Rahmen fokussiert. Dem liegt die oben beschriebene Berechnung der Position in der Bilddatei zugrunde. Nun können die Melioratoren das Buchstabenbild mit der OCR-Zeichenkette mit geringer Mühe vergleichen.

Dies geht aus der folgenden Abbildung hervor.

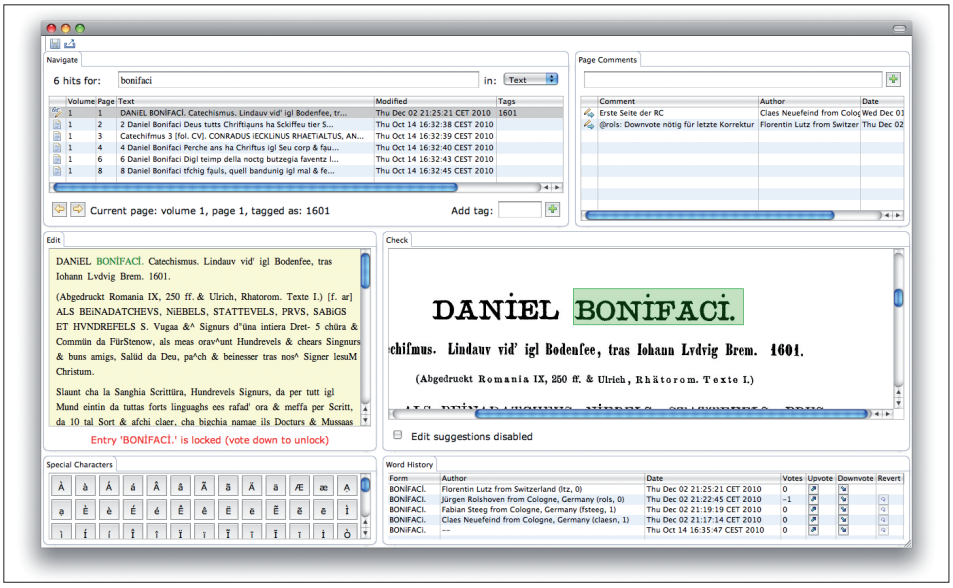


Fig. 11: Editorwerkzeug: im linken, gelb unterlegten Text wurde BONIFACI (grüne Schrift) angeklickt. Aus den Positionsinformationen wurde der grün unterlegte Rahmen im mittleren rechten Fenster in der Bilddatei berechnet.

In der Abbildung des Editorwerkzeugs findet sich im linken, gelb unterlegten Teilfenster der OCR-gelesene Text. Rechts sieht man das Ausgangsimage. Im linken Teilfenster wurde das Wort BONIFACI angeklickt und erscheint deshalb in grüner Schrift. Im rechten Fenster mit dem Image ist das Buchstabenbild BONIFACI durch einen grünen Rahmen unterlegt (hier sei nochmals ausdrücklich darauf verwiesen, dass das rechte Fenster keinen Text mit einzelnen, diskreten Buchstaben enthält, sondern lediglich ein Bild, quasi eine Sammlung von schwarzen, weißen oder grauen Punkten). Dies unterstützt die Arbeit der Melioratoren: Für Korrekturen braucht man sich nicht auf ihre Sprachkompetenz alleine zu verlassen, sondern man ist stets an die textuelle Grundlage (in Gestalt der Images aus den digitalisierten Seiten der Chrestomathie) gekoppelt.

Damit stellt sich die Frage, wie die Verbindung zwischen der OCR-gelesenen Zeichenkette links und dem Buchstabenbild rechts hergestellt wird. Der ABBYY-*Finereader* notiert in einem recht speziellen PDF-Format zu jedem erkannten graphischen Wort dessen Position im Ausgangsimage. Dabei werden vier Koordinaten gespeichert. Sie spiegeln die Ecken des grünen Rechtecks wider. Die Extraktion der Wörter mitsamt ihrer Positionskoordinaten erfolgte mit der Software-Bibliothek PDFBox (<http://pdfbox.apache.org/>). Die ausgelesenen Informationen

(Wort und Position) werden in XML-Form abgelegt und stellen die Grundlage für das Highlighting-Feature in der Korrekturumgebung dar. In dieser sind die Koordinatenwerte natürlich wechselnden Fenstergrößen anzupassen.

Die Arbeit der Melioratoren wird auch dadurch unterstützt, dass alle Vorkommen eines graphischen Wortes über die Suchfunktion ermittelt werden können. Die obige Abbildung zeigt im linken oberen Fenster die Vorkommen von BONIFACI in der Chrestomathie. Das ist insofern auch hilfreich, weil es bei der OCR auch systematische Fehler gibt. Systematische Fehler eines Worttypen *(types)* erstrecken sich über alle Vorkommen *(tokens)*. Solche Fehler lassen sich auf diese Weise rasch verbessern.

Das rechte obere Fenster hat die Funktion, Kommentare und Auszeichnungen darzustellen. In der finalen Version werden sich die Melioratoren in diesem Fenster über ihre Arbeit austauschen und die anreichernden Annotationen einbringen.

Das linke untere Fenster ermöglicht es, in den OCR-gelesenen Text Zeichen einzufügen, die auf normalen Tastaturen und in den gängigen Zeichensätzen nicht enthalten sind. Solche Zeichen wurden gerade in den ältesten gedruckten Texten des Bündnerromanischen häufig gebraucht.

Das Editorwerkzeug findet in einem kollaborativen System Verwendung. Dies hat zur Folge, dass alle Melioratoren die Texte verändern können. Allerdings ist potentiell nicht jede Veränderung auch eine Verbesserung. Das ist freilich kein singuläres Problem des DRC-Projekts, sondern ein generisches aller kollaborativen Systeme. Hier liegt also eine Lösung nahe, wie sie mit Erfolg etwa von der Wikipedia praktiziert wird: Es werden alle Versionen der Veränderung abgespeichert. Daraus ergibt sich auch die Möglichkeit, das System in einen früheren Zustand zurückzuversetzen. Für das DRC-Projekt wird dieses Prinzip in dem Sinne umgesetzt, dass für jedes graphische Wort eine Liste angelegt wird, in der eine neue, verbesserte Form abgespeichert wird. Eine solche Liste repräsentiert eine Art Verbesserungs- oder Veränderungsgeschichte; an ihr lässt sich auch beobachten, wie die Melioratoren mit den Texten umgehen.

Das rechte, untere Fenster zeigt dieses Konzept beispielhaft für das angeklickte Wort BONIFACI. Es führt die Wortgeschichte von BONIFACI auf, es nennt die Namen derer, die an der Wortgeschichte gearbeitet haben, und es skizziert ein Belohnungs- und Wettbewerbssystem, das über die Tätigkeit der Melioratoren Buch führt (im Hinblick auf Datenschutz: Anonyme Melioratoren sind zugelassen, die Teilnahme am Bewertungssystem ist freiwillig).

Gerade im Hinblick auf Minderheitensprachen sehen wir in einem solchen System ein Instrument zur interaktiven Teilhabe und Kommunikation. Daraus lässt sich ein Benutzerverhalten ablesen, das nicht nur Auskunft über Schwächen des Systems (und somit über Optimierungsbedarf) gibt, sondern das auch Sprachverhalten, Sprachbewusstsein, Wissen über Sprache und ihren Gebrauch, auch ihren Wandel abbildet.

Dies könnte die Grundlage für sehr unterschiedlichen Einsatz sein. Eine der Ideen war, ein solches System im Rahmen von Schulprojekten einzusetzen.[17]

Ein interaktives soziales System birgt die Gefahr, dass Verbesserungsvorschläge nicht akzeptiert werden, dass sie wiederholt vorgenommen und wiederholt zurückgewiesen werden. Bei der Wikipedia spricht man in diesem Zusammenhang von *edit wars*.

Im DRC-Projekt werden potentielle Konflikte dieser Art durch den Moderator, einen ausgewiesenen Fachmann und Muttersprachler, gelöst.

Dies spiegelt auch eine grundsätzliche Aufteilung des Projekts wider. Im Projekt kooperieren zwei Gruppen: die Computerlinguisten in Köln und die vorwiegend in Bünden operierende Gruppe der Melioratoren und der Moderator. Dadurch werden nicht nur die jeweiligen fachlichen Stärken zusammengeführt, sondern auch die Autonomie und Selbstbestimmung der Sprecher der Minderheitensprache gewahrt und gestärkt. Das Romanische bleibt den Romanen.

Ein primäres Anliegen des Projekts ist die Erstellung korrekter Texte. Korrektheit heißt hier: der Vorlage der gedruckten Chrestomathie entsprechend.[18] Dabei ist es sinnvoll, möglichst viele Fehler automatisch oder halbautomatisch zu verbessern. Das Editorwerkzeug unterstützt unterschiedliche Formen der Fehlerkorrektur. Eine Form setzt auf Korrekturlexika und schlägt bei abweichenden Vorkommen im Text Alternativen vor. Bei diesem Vorgehen wird die Editierdistanz zwischen den Vorkommen in Text und Lexikon berechnet.[19] Voraussetzung ist freilich ein Lexikon. Allerdings kann ein Lexikon nur sinnvoll für ein Idiom eingesetzt werden, wenn der Text orthographisch normiert ist. Letzteres ist bei älteren Texten nicht der Fall, mehr noch, Texte sind – etwa im Fall der Bifrun-Bibel – mitunter inkonsistent verschriftlicht.

[17] Persönliche Mitteilung von Esther Krättli.

[18] Dabei stellt sich aber auch die Frage, wie mit offensichtlichen Druckfehlern der Vorlage umzugehen ist. Diese können kommentiert werden.

[19] Levenshtein-Distanz, cf. Gusfield 1997, 216.

Für die neueren Texte der verschiedenen Idiome gibt es jeweils (auch maschinenlesbare) Lexika. Für die älteren Texte ist ein solches Lexikon – besser: eine Wortliste – zu erstellen. Pragmatisch bietet sich dafür der folgende Weg an: Mit Hilfe der Metadaten wird der jeweilige Text nach diatopischen und diachronen Bedingungen seligiert; aus bereits korrigierten Teilen des Textes wird eine Wortliste für das Korrekturlexikon erzeugt. Die Verwendung manuell (d.h. durch die Nutzer) korrigierter Formen gewährleistet, dass nur relevante Korrekturvorschläge gemacht werden.

Eine weitere Korrekturform operiert mit regulären Ausdrücken. Eine bekannte Form regulärer Ausdrücke findet sich in den meisten Textverarbeitungsprogrammen bei der Suche von Zeichenketten. Sucht man alle Wörter, die mit "in-" beginnen, so schreibt man "in*". Reguläre Ausdrücke höherer Komplexität erlauben eine sehr flexible Suche und auch ein automatisches oder halbautomatisches Ersetzen.

Die Typographie der ersten Bände der Chrestomathie verwendet ein langes Fraktur -S, das von der OCR meist nicht von einem f unterschieden werden kann (der einzige Unterschied ist der waagerechte Strich im f). Dies zeigt der folgende *Keyword in Context*-Ausschnitt.

035 - 008 42	i ftudij. abbeverare le beftie. A. A. avunda.	**Abadeffa**	Avat. bandunar. darsi il ftudi. boventâ. Romancio di
036 - 026 75	Per viver et morir beada. 175 Pro temp regueva l	**Abadefsa**	Adelheit, nativa da Novena, Prudainta et pia contefsa
011 - 021 15	ton de quescher E ton de surportar. 5 La mumma 1'	**abadessa**	Ha in fasierli tgau; Ella scogna avon messa E sco ch'
012 - 016 59	las giias stouvan strasuner, 25 E Par nus tingner	**abadientt**	alg fat CunPlieu fün alg mendar TimP, schi
030 - 019 41	enten Jesum Christum, udit ilg plaid da Deus,	**abadit**	à quell, milgureit vofsas vittas, vus parrieit da 40 morir, vivit
037 - 029 17	Ilg Mastral da Criminal dürond feis offici non s def	**abafar**	de gradu ad ir cun (s h) raanadüras da dor punt Martina
012 - 027 11	Pg afchaid, fchi dis el: Elg es cumplieu: & hauiand	**abaffò**	Pg chio fchi det 10 el fù Pg fpiert. Et Ps ludeaus par ch' elg
012 - 029 40	è lg fquitfchad. Deis ils fuperwis fpltimmel,	**Abaffa**	lg adutzad. Parchê chia tti inuidafch, Meis Deis chi nun m'
017 - 066 15	dal utfchê, è tuot las chantadtiras veguen à gnir	**abaffadas**	7. Et eir il craftiaun vain à tmair dals louhs sti aut, ed
017 - 047 35	ilg fo3 dalg infierne. 112. Cun tuot avaunt Deis ns'	**abaffain**	Da eottr, chi ais grond tempe, 780 Et nofs puehiads tuots
017 - 055 37	E lur fadifs ils han fquitfchads, & fut lur man fun eus	**abaffats**	43. Bleras votas ils ha el deliberats; ma [p. 309] eus il
024 - 007 14	fch' els fuoffen informâts il nöbel ftadi del Baur non	**abaffeSen**	quel uSchea pro tuottas natziuns las plü ilüminâdas
017 - 013 40	giaglard in la mtidaeda cun la quaela tti vainft	**abaffo**	Perche fco Pg òr cun l'g foe vain appruo, vfchea la lieud
037 - 032 5	lugüu Laudama dürontt feis officij non defs tant	**abafsar**	a far il fuorman fiuclegia or da nofs pajais tant dominant
035 - 008 33	Romancio di Surfelva. Italiano. A. A (Propofitione)	**abaftanza**	abbadezza. abbate. abbandonare. abbandonar i ftudii.

Fig. 12: *Keyword in Context* - Types

Viele der Lesefehler lassen sich leicht korrigieren, vor dem Hintergrund, dass es im Bündnerromanischen keine (oder wohl kaum) Kombinationen von "f" und "s" oder von "f" und "t" gibt. Für andere Fälle bietet sich ein gegebenenfalls gewichteter Levenshtein-Algorithmus an.[20]

Melioration ist Korrektur und Anreicherung. Anreicherung ist die Auszeichnung der Texte mit Metadaten. Ein Teil der Metadaten kann übernommen werden

[20] Cf. erneut Gusfield 1997, 216.

(aus dem Projekt Digizeitschriften), andere können automatisch durch die Auswertung des ausgezeichnet gestalteten Registerbandes von EGLOFF und MATHIEU erzeugt werden (dazu müssen die Eintragungen des Registerbandes erst OCR-erfasst und dann geparst werden).

Die dritte Quelle für Metadaten ist das Wissen der Melioratorengemeinschaft. Die Nutzung dieser Quelle eröffnet neue Möglichkeiten für den Umgang und die Rezeption von Texten. In traditioneller Sicht sind Texte unveränderbare Objekte; hier jedoch interagieren sie mit der Gemeinschaft, die sich über das Netz konstituiert. Durch den Umgang mit der Gemeinschaft verändern sie sich; es ist die Gemeinschaft, die den Veränderungsprozess steuert. Texte werden nicht mehr alleine konsumiert, sie sind nicht mehr *read only*.

Natürlich muss Sorge dafür getragen werden, dass der objektive Charakter der textuellen Basis erhalten bleibt, sozusagen der kritische Text. Dies ist jedoch leicht vereinbar, wenn die textuelle Basis und die Auszeichnungen getrennt bleiben (dieses Prinzip ist übrigens im Hinblick auf Texteditoren sehr klar in dem Beitrag von GUTKNECHT 1985 dargestellt).

Ein weiterer Vorteil des Community-basierten Ansatzes mit der Dokumentation von Korrekturen und Anreicherungen ist der Lerneffekt, der sich aus der Orientierung am Benutzerverhalten und an der besten Praxis ergibt. Schließlich können auch gewisse Wettbewerbseffekte erzielt werden, wenn Bewertungsranglisten erstellt werden.

7. Implementation

Das gesamte System beruht auf *Java*-Technologien. Damit profitiert das Vorhaben von modernen, weitestgehend portablen und quelloffenen Standards. Ein Teil der Software ist in *Java* geschrieben, ein anderer in *Scala*. Die gemeinsame Basis ist *Java-Byte-Code*, in den compiliert wird. Mit *Java-Byte-Code* als verbindender Basis bleibt das System zukunftsoffen (es gibt mehr als 200 to-*Java-Byte-Code Compiler*).

Die folgende Skizze gibt einen Überblick über die Systemarchitektur.

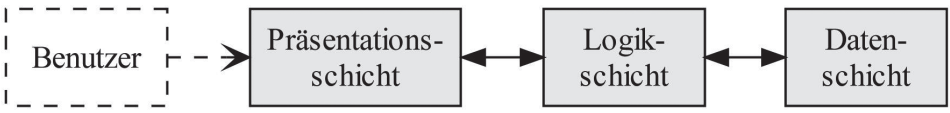


Fig. 13: Architektur

Das System ist in klassischer dreischichtiger Architektur aufgebaut, in der der Benutzer über die Präsentationschicht – die graphische Oberfläche – mit der Verarbeitungsschicht, der sogenannten Logikschicht, verbunden ist. Die Programmlogik legt ihre Daten in der Datenschicht ab bzw. bezieht sie von dort. Dabei handelt es sich im wesentlichen um die OCR-gelesenen Texte, die darin vorgenommenen Verbesserungen, die Images der Digitalisierung und die Positionen in den Images, die den Melioratoren die Verbindung zwischen Text und Image zur Anzeige des angeklickten Wortes herstellen. Zusätzlich enthält die Datenschicht Auszeichnungen bzw. Tags sowie die Kommentare (zur Anreicherung der Texte).

Diese grundlegende Architektur wird schrittweise weiterentwickelt. Ausgangspunkt ist dabei die erste funktionale Variante ("Beta 1").

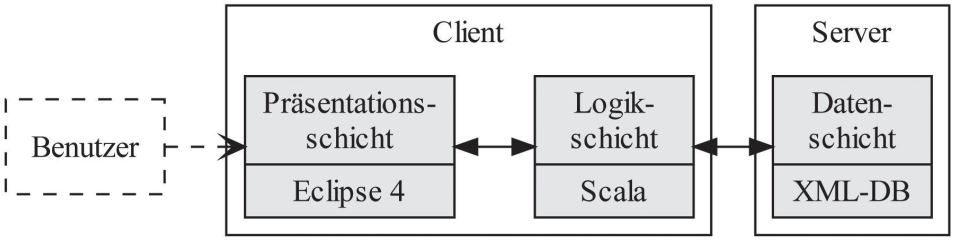


Fig. 14: Beta 1

In dieser ersten funktionalen Variante werden die Daten in einer XML-Datenbank serverseitig vorgehalten, während die Ausführungs- und die Präsentationsschicht clientseitig realisiert werden. Dies führt zu einem vergleichsweise mächtigen (einem "Fat-") Client.

Die weiteren Entwicklungsschritte werden den Client abspecken (zu einem "Thin Client"). Die Logikschicht wird dann zum Server verlagert.

Im nächsten Schritt kann über eine Portalseite auch direkt auf die textuellen Daten zugegriffen werden. Der Zugriff auf diese Daten wird durch Suchfunktionen erleichtert.

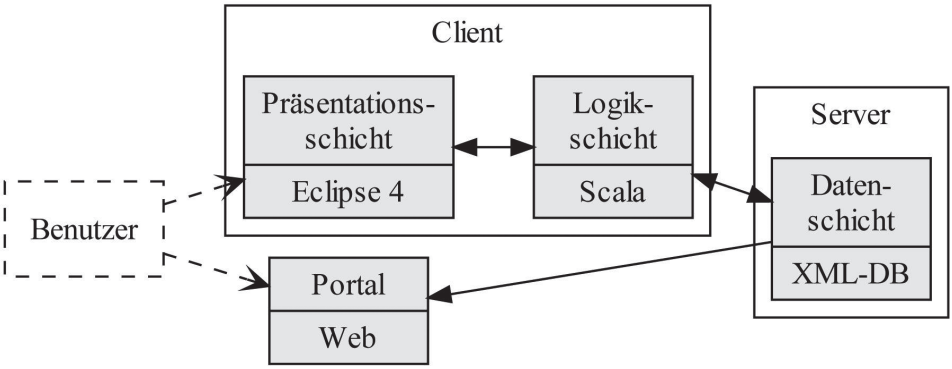


Fig. 15: Beta 2

Die obige Skizze spiegelt den Stand der Entwicklung im März 2011 wider.

Final wird der Zugriff über eine Web-Oberfläche mit RAP-Technologien[21] erfolgen. Alternativ wird eine Offline-Variante angeboten, die der bisherigen Lösung sehr nahe bleibt. Hintergrund dieser zweigleisigen Architektur ist der Wunsch, das System auch dann einsetzen zu können, wenn es keine Netzverbindungen gibt. Die Kommunikationsinfrastruktur der Bündnerromanen ist nicht überall in der Welt vorhanden; gleichwohl sollte – im Sinne einer "Repatriation", d.h. dem Zurückbringen von Texten, die etwa in den Bibliotheken ehemaliger Kolonialmächte aufbewahrt werden – das System eingesetzt werden können (cf. Kap. 8, Potentiale).

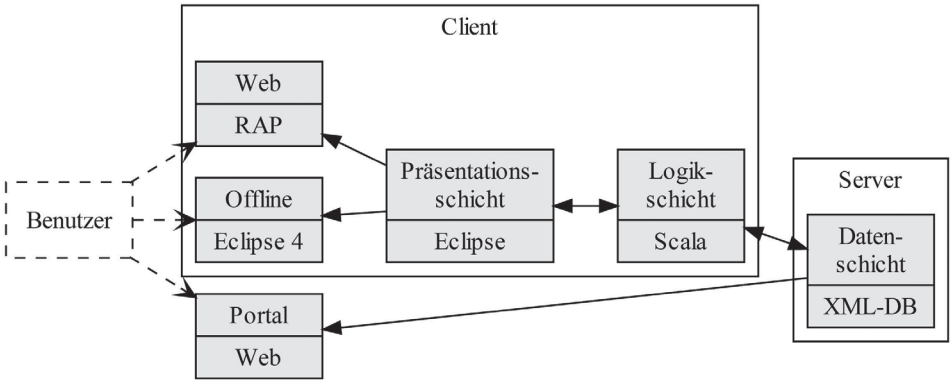
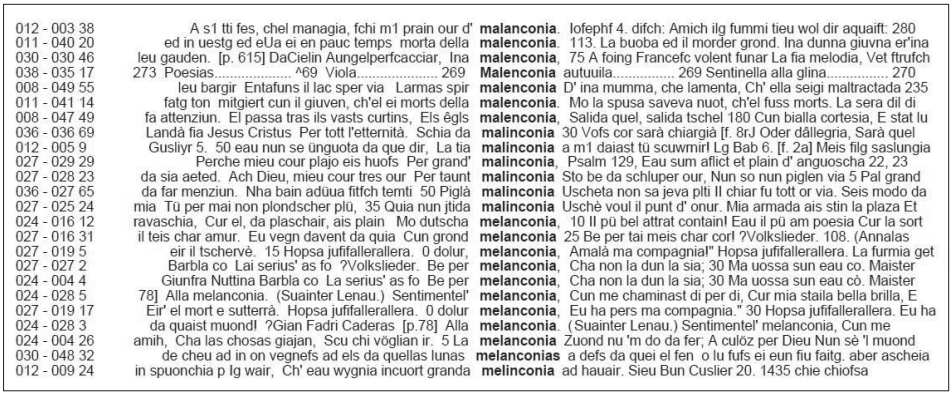


Fig. 16: Komplettsystem

[21] *Rich Ajax Platform*, cf. <http://www.eclipse.org/rap/>.

8. Potentiale

Zu den einfacheren Anwendungen mit maschinenlesbaren Texten zählen Formen der Suche, die beispielsweise mit Hilfe regulärer Ausdrücke umgesetzt werden können. Reguläre Ausdrücke ermöglichen es, Suchen flexibel zu gestalten. Dafür werden Suchmuster angegeben, bei denen an gewissen Positionen nicht ein einziger, sondern unterschiedliche Zeichen vorkommen können, z.B. ein "a" oder ein "e". Auch können Zeichen ausgelassen oder übersprungen werden. Kombiniert man diese und weitere Variationen in einen regulären Ausdruck, so kann man ein Netz aufspannen, in dem ähnliche Formen eingefangen werden. Dies zeigen die nachfolgenden Formen von "melancholia".[22]



012 - 003 38	A s1 tti fes, chel managia, fchi m1 prain our d'	**malanconia**	Iofephf 4. difch: Amich ilg fummi tieu wol dir aquaift: 280
011 - 040 20	ed in uestg ed eUa ei en pauc temps morta della	**malenconia**	113. La buoba ed il morder grond. Ina dunna giuvna er'ina
030 - 030 46	leu gauden. [p. 615] DaCielin Aungelperfcacciar, Ina	**malenconia**	75 A foing Francefc volent funar La fia melodia, Vet ftrufch
038 - 035 17	273 Poesias..... ^69 Viola..... 269	**Malenconia**	autuuila..... 269 Sentinella alla glina..... 270
008 - 049 55	leu bargir Entafuns il lac sper via Larmas spir	**malenconia**	D' ina mumma, che lamenta, Ch' ella seigi maltractada 235
011 - 041 14	fatg ton mitgiert cun il giuven, ch'el ei morts della	**malenconia**	Mo la spusa saveva nuot, ch'el fuss morts. La sera dil di
008 - 047 49	fa attenziun. El passa tras ils vasts curtins, Els êgls	**malenconia**	Salida quel, salida tschel 180 Cun bialla cortesia, E stat lu
036 - 036 69	Landà fia Jesus Cristus Per tott l'etternità. Schia da	**malinconia**	30 Vofs cor sarà chiargià [f. 8rJ Oder dållegria, Sarà quel
012 - 005 9	Gusliyr 5. 50 eau nun se ünguota da que dir, La tia	**malinconia**	a m1 daiast tü scuwmir! Lg Bab 6. [f. 2a] Meis filg saslungia
027 - 029 29	Perche mieu cour plajo eis huofs Per grand'	**malinconia**	Psalm 129, Eau sum aflict et plain d' anguoscha 22, 23
027 - 028 23	da sia aeted. Ach Dieu, mieu cour tres our Per taunt	**malinconia**	Sto be da schluper our, Nun so nun piglen via 5 Pal grand
036 - 027 65	da far menziun. Nha bain adüua fitfch temti 50 Piglà	**malinconia**	Uscheta non sa jeva plti Il chiar fu tott or via. Seis modo da
027 - 025 24	mia Tü per mai non plondscher plü, 35 Quia nun jtida	**malinconia**	Uschè voul il punt d' onur. Mia armada ais stin la plaza Et
024 - 016 12	ravaschia, Cur el, da plaschair, ais plain Mo dutscha	**melanconia**	10 Il pü bel attrat contain! Eau il pü am poesia Cur la sort
027 - 016 31	il teis char amur. Eu vegn davent da quia Cun grond	**melanconia**	25 Be per tai meis char cor! ?Volkslieder. 108. (Annalas
027 - 019 5	eir il tschervè. 15 Hopsa jufifallerallera. 0 dolur,	**melanconia**	Amalà ma compagnia!" Hopsa jufifallerallera. La furmia get
027 - 027 2	Barbla co Lai serius' as fo ?Volkslieder. Be per	**melanconia**	Cha non la dun la sia; 30 Ma uossa sun eau co. Maister
024 - 004 4	Giunfra Nuttina Barbla co La serius' as fo Be per	**melanconia**	Cha non la dun la sia; 30 Ma uossa sun eau cò. Maister
024 - 028 5	78] Alla melanconia. (Suainter Lenau.) Sentimentel'	**melanconia**	Cun me chaminast di per di, Cur mia staila bella brilla, E
027 - 019 17	Eir' el mort e sutterrà. Hopsa jufifallerallera. 0 dolur	**melanconia**	Eu ha pers ma compagnia." 30 Hopsa jufifallerallera. Eu ha
024 - 028 3	da quaist muond! ?Gian Fadri Caderas [p.78] Alla	**melanconia**	(Suainter Lenau.) Sentimentel' melanconia, Cun me
024 - 004 26	amih, Cha las chosas giajan, Scu chi vöglian ir. 5 La	**melanconia**	Zuond nu 'm do da fer; A culöz per Dieu Nun sè 'l muond
030 - 048 32	de cheu ad in on vegnefs ad els da quellas lunas	**melanconias**	a defs da quei el fen o lu fufs ei eun fiu faitg. aber ascheia
012 - 009 24	in spuonchia p lg wair, Ch' eau wygnia incuort granda	**melinconia**	ad hauair. Sieu Bun Cuslier 20. 1435 chie chiofsa

Fig. 17: *Keyword in Context* im Zusammenhang mit einer Suche durch reguläre Ausdrücke.

Für die lexikographische und lexikologische Arbeit dürfte die Angabe des Alters der Texte von Bedeutung sein, gerade auch für die Dokumentation von Erstbelegen.[23]

Das nächste Beispiel zielt auf eine syntaktische Fragestellung. Das Surselvische zeichnet sich bei den Reflexiva dadurch aus, dass es für alle Personen eine einheitliche Form verwendet, auf die Erstellung einer Konkordanz mit der Person also verzichtet:

jeu selavel	"ich wasche mich"
ti selvas	"du wäschst dich"
el selava	"er wäscht sich"

[22] Diese beiden Suchen wurden in Absprache mit Matthias Grünert für die Arbeit an entsprechenden Lexikonartikeln des *Dicziunari Rumantsch Grischun* durchgeführt.

[23] Persönlicher Hinweis von Matthias Grünert.

Weiterhin findet sich das Reflexivum stets vor dem Bezugsverb, so jedenfalls auch z.B. Spescha 1989, 383–390[24], Beispiel *jeu sundel selavaus* "ich habe mich gewaschen".

Allerdings finden sich in der Chrestomathie folgende Beispiele bei dem Verb *laschar*:

Denton tut che selai *splenar, cura ch' ils sauns assistan als malsaus* [sic]... "Indessen lässt sich alles ausgleichen, wenn die Gesunden den Kranken beistehen, wenn die Reichen den Armen helfen..." (Decurtins 1982–1986, Bd. I, 640)[25]

Orda nossa patri'alpina mai selais ti bandischar. "Aus unserer Alpenheimat lässt Du dich nie verbannen." (Decurtins 1982–1986, Bd. XII, 115)

Hier handelt es sich um einen Spezialfall für Anhebung von Pronomina – in der Terminologie der generativen Grammatik als *Clitic Climbing* bezeichnet –, der im Surselvischen nur in sehr speziellen Domänen auftritt, nämlich bei faktitiven Verben.[26]

Über die gegebene Textbasis hinaus weist das Chrestomathieprojekt folgende Erweiterungen und Erweiterungspotentiale auf: Es verknüpft (Teile der) Chrestomathie mit Übersetzungen. Bereits jetzt stehen die Übersetzungen von Ursula Brunold-Bigler und Konrad Widmer dank des großzügigen Entgegenkommens der Autoren und des Verlags zur Verfügung.[27] Daraus ergeben sich für die maschinelle Weiterverarbeitung neue Impulse, etwa zur automatischen Erstellung von Gebrauchslexika aus Paralleltexten.[28]

[24] Es sei darauf hingewiesen, dass sich die bündnerromanischen Idiome im Bezug auf die Objektklitika sehr unterschiedlich entwickelt haben; das Engadinische folgt dem romanischen Typ.

[25] Hier handelt es sich um einen Druckfehler. Das Riesenwerk der Chrestomathie enthält verständlicherweise einige Fehler dieser Art. Für die Korrektur stellt sich die Frage, ob über Lesefehler hinaus auch Druckfehler verbessert werden sollen; dies würde das Prinzip der vollständigen Wiedergabe der Vorlage verletzen. Genau hier empfiehlt es sich, die Textstelle durch Kommentierung anzureichern.

[26] Diese Verbklasse subkategorisiert möglicherweise direkt ein Verb, nicht jedoch – wie andere übergeordnete Verben ("Matrixverben") – sogenannte funktionale Verbalkategorien wie TP (Tempusphrase). Daher ist nach einem Verb wie *laschar* auch die Negation der eingebetteten Phrase ausgeschlossen (cf. dazu auch die Beobachtung, dass in romanischen Sprachen nach faktitiven Verben keine Negation auftreten kann; persönliche Mitteilung Guido Mensching, Berlin). Ich begnüge mich mit diesen Hinweisen. Ich bin der Ansicht, dass die gerade im Bündnerromanischen heterogene Syntax der Pronomina auch von areallinguistischem Interesse ist und dass eine digitale Textsammlung datengetriebene Forschung wirksam unterstützen kann.

[27] Cf. Brunold-Bigler/Decurtins 2002 und Brunold-Bigler/Widmer 2004. Den genannten Autoren und dem Casanova Verlag, Herrn Bühler, Chur, sei hier sehr herzlich für die großartige Unterstützung gedankt.

[28] Cf. Martin Kay <http://www.stanford.edu/~mjkay/CYCLING.pdf>.

Den Wiki-Prinzipien des Projekts folgend können Übersetzungen grundsätzlich auch von interessierten und kompetenten Einzelpersonen verfasst werden. Damit wird die Rätoromanische Chrestomathie inhaltlich auch jenen zugänglich, die das Bündnerromanische nicht beherrschen.

Die Rätoromanische Chrestomathie wird durch den Einsatz der Melioratoren in einem interaktiven Prozess beträchtlich erweitert und qualitativ aufgewertet. Diese Aufwertung kann durch eine Georeferenzierung der in den Texten enthaltenen Toponyme gesteigert werden. Georeferenzierung ist mehr als nur eine Verknüpfung von Texten und Sprache mit der außersprachlichen Welt; sie ermöglicht vielmehr neue Gebrauchsweisen, z.B. durch Präsentation und Nutzung der Texte auf mobilen Geräten, die z.B. dem Wanderer die Texte anzeigt, die die erwanderte Region toponymisch benennen. Als Beispiel sei das Lugnezer Frauentor (romanisch *Porclas*) gewählt – dort, wo die tapferen Lugnezerinnen das eindringenden Heer des Grafen Rudolf von Montfort 1355 abwehrten. Ein Bezug zu diesem Ort findet sich in der Chrestomathie in den Volksliedern:

> *Àch co mi fa mol il cor, Che miu muronz ei jus da Porclas or.* "Ach wie schmerzt es mein Herz, dass mein Schatz über das Frauentor herausgegangen ist."[29]

Fig. 18a: Text und Georeferenzierung: *Porclas* (Das Frauentor), Lugnez

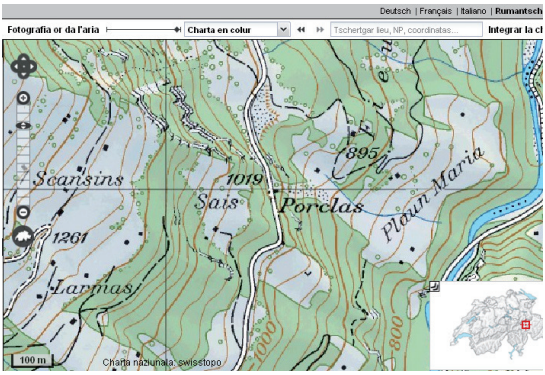


Fig. 18b: Text und Georeferenzierung: Karte Lugnez

[29] Decurtins 1982–1986, Bd. II und III, 294.

Abschließend sei auf eine weitere Möglichkeit der Verbindung von Geoinformationen und den Texten der Chrestomathie verwiesen. Das Bündnerromanische beschreibt Positionen und Positionsveränderungen in der gebirgigen Umwelt außerordentlich präzise; darauf hat in einem wegweisenden Aufsatz Helmut Lüdtke (1953) verwiesen; die Thematik ist auch von Theodor Ebneter (1984) und rezent von Raphael Berthele (2006) aufgegriffen worden. Für weiterführende Untersuchungen dieser Art dürfte die Digitale Rätoromanische Chrestomathie mit Zugriff auf Geoinformationen wie Höhe und Ausdehnung ein ausgezeichnetes Hilfsmittel sein.

Aber auch editionsphilologisch würde die Digitale Rätoromanische Chrestomathie eine gute Grundlage für modernes textkritisches Vorgehen bieten.

Somit ist die Digitale Rätoromanische Chrestomathie nicht mehr länger ein statischer, abgeschlossener Text, sondern ein facettenreicher, offener, interaktiver Prozess, eine Kommunikation zwischen Computerlinguisten und Sprechern, ein Prozess der Aneignung und Selbstermächtigung[30] der Sprecher, eine Rückeroberung der Sprache und des kulturellen Erbes durch ein neues, smartes und interaktives soziales Werkzeug. Hier geht es somit nicht oder nicht nur um Musealisierung und Folklorisierung.[31] Die Digitale Rätoromanische Chrestomathie stärkt die Sprache durch Auseinandersetzung mit der Sprache.

Auch der Sprachgegenstand ist nicht auf die Texte der Chrestomathie beschränkt. Vielmehr sollte die Chrestomathie Ausgangs- und Kristallisationspunkt für eine weitergehende, vielleicht sogar vollständige Digitalisierung des Bündnerromanischen sein. Erweiterungen der Textbasis über die Chrestomathie hinaus sind in ersten Ansätzen durch die Digitalisierung der *Tschantamaints* (Schorta 1982a und 1982b) und der vollständigen Bifrun-Bibel in der Edition von Gartner (1913) gemacht.[32] Ein Vorgehen *ab ovo* – i.e. ausgehend von den Bibelübersetzungen des Bündnerromanischen – ist eine lohnende Zukunftsaufgabe.

[30] Terminus von Claes Neuefeind (cf. Neuefeind/Steeg 2011, im Druck).

[31] Vielleicht stellt eine Georeferenzierung von Chrestomathietexten in mobilen Geräten eine neue Form der Folklorisierung dar; wir glauben jedoch, dass die Kombination von modernen Technologien und sozusagen "uralten" Sprachen (es gehört wohl zum Mythos von Kleinsprachen, diese seien uralt, vergleiche etwa Giachen Caspar Muoths Aufruf an die Romanen "*defenda, Romonsch* [sic!]*, tiu vegl lungatg!*", Decurtins 1982–1986, Bd. 1, 676) letztere aktiv schützt.

[32] Cf. auch Bd. 5 der Chrestomathie, p. 251 in der Octopus-Ausgabe.

9. Eine Modell für Minderheitensprachen?

Während der Projektlaufzeit liefen zahlreiche Fragen ein, ob es nicht möglich sei, das für die Chrestomathie geschaffene System für andere Sprachen – vorwiegend Minderheitensprachen – einzusetzen. In der Tat hat die Digitale Rätoromanische Chrestomathie prototypischen Charakter. Im romanischen Kontext würden sich Sprachen wie das Aromunische oder das Sardische für Projekte mit ähnlicher Zielsetzung anbieten. Ein weiteres interessantes Anwendungsfeld wären Kreolsprachen; das hier vorgestellte System würde die geographisch weit getrennten Sprachen über das Web verbinden und die sprachlichen und wissenschaftlichen Communities zusammenführen. Von besonderer Bedeutung ist hierbei die spezielle Eigenschaft des Systems, sowohl online als auch offline arbeiten zu können.

Häufig finden sich Texte und Textsammlungen in den Bibliotheken ehemaliger Kolonialmächte, sind aber in den Herkunftsregionen nicht verfügbar. Diese Texte können in digitaler Form den Sprachgemeinschaften der Herkunftsregionen zurückgegeben und von ihnen korrigiert und angereichert werden – auch offline, wenn kein direkter Netzzugang besteht.

10. Danksagung

Das Projekt der Digitalen Rätoromanischen Chrestomathie wäre ohne die Förderung in Deutschland und in der Schweiz nicht zu verwirklichen gewesen. Unser Dank gilt daher der Deutschen Forschungsgemeinschaft, die das Projekt im Rahmen des Förderbereichs "Wissenschaftliche Literaturversorgungs- und Informationssysteme" zwei Jahre lang, von November 2009 bis Oktober 2011 unterstützt hat.

In der Schweiz fand das Projekt im Legat Anton Cadonau, im Institut für Kulturforschung Graubünden und im Amt für Kulturförderung des Kantons Graubünden höchst willkommene finanzielle Unterstützung. Weitere Förderung erhielt das Projekt durch die *Societad Retoromontscha*.

All diesen Einrichtungen schulden wir unseren herzlichsten Dank.

Die studentischen Mitarbeiter Frauke Schmidt und Michail Atanassov der Sprachlichen Informationsverarbeitung tragen durch ihr Wissen, Können und Engage-

ment sehr zur erfolgreichen Durchführung des DRC-Projekts bei. Unser Dank geht auch an Herrn Michele Badilatti für den Einsatz bei der Melioration der digitalen Chrestomathie und für die vielen Hinweise zur Verbesserung der frühen Softwareversionen.

11. Bibliografie

Berthele, Raphael: *Ort und Weg. Die sprachliche Raumreferenz in Varietäten des Deutschen, Rätoromanischen und Französischen*, Berlin/New York 2006.

Brunold-Bigler, Ursula / Decurtins, Caspar: *Die drei Winde. Rätoromanische Märchen aus der Surselva,* Chur 2002.

Brunhold-Bigler, Ursula / Widmer, Konrad: *Die drei Hunde. Rätoromanische Märchen aus dem Engadin, Oberhalbstein, Schams,* Chur 2004.

Decurtins, Caspar: *Rätoromanische Chrestomathie*, Erlangen 1888–1919, 13 voll.; [als Faksimile mit Register, in 15 voll. neu hgg. von Octopus-Verlag / Società Retorumantscha, Chur 1982–1986].

Deplazes, Gion: *Dus plaids sin via*, in: Egloff/Mathieu 1986, op.cit., 5–6.

Ebneter, Theodor: *Die Adverbien und Präpositionen des Ortes und der Richtung im Romanischen von Vaz/Obervaz,* in: "Zeitschrift für romanische Philologie", 100/3–4, 1984, 387–407.

Egloff, Peter / Mathieu, Jon: *Rätoromanische Chrestomathie. Begründet von Caspar Decurtins*, Bd. 15: Register, Chur 1986.

Gartner, Theordor: *Das Neue Testament. Erste Rätoromanische Übersetzung von Jakob Bifrun 1560*, Halle/Saale 1913.

Gusfield, Dan: *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology,* Cambridge 1997.

Gutknecht, Jürg: *Concepts of the Text Editor Lara*, in: "Communications of the Association for Computing Machinery", 28, 9, 1985, 942–960.

Liver, Ricarda: *Rätoromanisch. Eine Einführung in das Bündnerromanische*, Tübingen 1999.

Lüdtke, Helmut: *Präpositionen der Orts-, Höhen- und Richtungsbezeichnung im Graubündner Oberland*, in: "Romanische Forschungen", 66, 1953, 374–378.

Mcenery, Tony / Wilson, Andrew: *Corpus Linguistics*, Edinburgh 2001^2.

Neuefeind, Claes / Steeg, Fabian: "*Stai si, defenda, Romontsch, tiu vegl lungatg" – Digitalisierung als Mittel kultureller Selbstermächtigung kleinerer Sprachgemeinschaften,* [im Druck].

Riatsch, Clà: *Mehrsprachigkeit und Sprachmischung in der neueren bündnerromanischen Literatur,* Chur 1998.

Schorta, Andrea: *Tschantamaints d'Engiadina bassa / Die Dorfordnungen des Unterengadins*, in: Rechtsquellen des Kantons Graubünden, Serie B: Dorfordnungen, Bd. 1, Schlarigna 1982^{2}a.

Schorta, Andrea: *Tschantamaints d'Engiadin'ota, da Bravuogn e Filisur / Die Dorfordnungen des Oberengadins, von Bergün und Filisur,* in: Rechtsquellen des Kantons Graubünden, Serie B: Dorfordnungen, Bd. 2, Schlarigna 1982^{2}b.

Spescha, Arnold: *Grammatica sursilvana*, Chur 1989.

Wigger, Bernard: *Die Schweizerische Konservative Volkspartei 1903–1918: Politik zwischen Kulturkampf und Klassenkampf,* Fribourg 1997.

Ressumé

Chest contribut prejenta n sistem de analisa digitala dla Rätoromanische Chrestomathie (RC) de Caspar Decurtins, publicheda tla revista "Romanische Forschungen" (Erlangen, 1896–1919). La RC é enchina encuei la recoiuda de tesć plu emportanta per l rumanc y na fontana de mascima emportanza per la linguistica, la leteratura y la etnologia. L'analisa digitala met a desposizion te na forma daverta *(open source)* duc i tesć dla RC per enrescides linguistiches y filologiches dl corpus. L medem vel ence per les metodes de miorazion y de corezion svilupedes tl cheder dl projet de enrescida. Tres n Wiki végnel metù adum la corezion automatica y interativa. La corezion interativa lieia la comunité linguistica retoromana. Les tecniches de chest projet dessa daldò gnì adoredes te d'autres recoiudes de tesć dl rumanc y de d'autri lingac. La scomenciadiva à enscila n carater prototipich per projec, che se dà ju con la analisa de recoiudes de tesć spezialisés per lingac de mendranza, ma ence per de tei che é en pericul.