Collaboratively Building Corpora - a Case Study for Romansh

Claes Neuefeind

University of Cologne, Department of Linguistics, Linguistic Data Processing
Albertus-Magnus-Platz, 50923 Köln, Germany
E-mail: neuefeind@spinfo.uni-koeln.de

Abstract. This paper presents an approach to collaboratively building corpora through digitisation. The discussion is based on the project *Digital Romansh Chrestomathy* (DRC)[1], which deals with the collaborative correction and enrichment of a Rhaeto-Romanic text collection using a web-based editing environment. As a side effect, the language community is integrated into the process of an active language preservation.

Keywords: Collaborative Correction, Collaborative Annotation, Digitisation, Historical Corpora, Cultural Heritage

1 Introduction

The DRC is part of a series of projects aiming at creating a Rhaeto-Romanic corpus based on an existing historical text collection. The DRC project focuses on the collaborative correction of recognition errors from Optical Character Recognition (OCR), comparable to the approaches of the *Australian Newspapers Digitisation Program* (ANDP)[2], which successfully established a web-based platform for correcting digitised newspapers (cf. [3]), and *Wikisource*[3], a sister project of Wikipedia dedicated to collaboratively building a library of source texts.

While the basic goal of providing a collaboratively corrected Rhaeto-Romanic text corpus can be claimed to be achieved, a number of methodological questions arose during the project. After briefly sketching our approach, these questions will be discussed with regard to the specific language situation for Romansh.

2 The Digital Romansh Chrestomathy

Text basis of the DRC project is the "Rätoromanische Chrestomathie" (Romansh Chrestomathy, RC) by Caspar Decurtins [1], originally published 1888-1919 in the journal "Romanische Forschungen" (Erlangen: Junge). With approx. 7500 pages, the RC can be regarded as the most important text collection of the Rhaeto-Romanic language (cf. [2]).

[1] http://www.crestomazia.ch
[2] http://www.nla.gov.au/ndp
[3] http://wikisource.org

Central to our approach is the collaborative editing environment[4] shown in figure 1, which juxtaposes the OCR texts and the corresponding digital images.[5] The pages can be selected according to the metadata adopted from the project *Digizeitschriften*. Following wiki-principles, corrections, annotations, and comments can be contributed by volunteers. All changes are logged (combined with information on the editing user and a time stamp), providing an editing history for every word token. Based on this, a simple rating system was introduced to keep track of the editing effort of the different users.

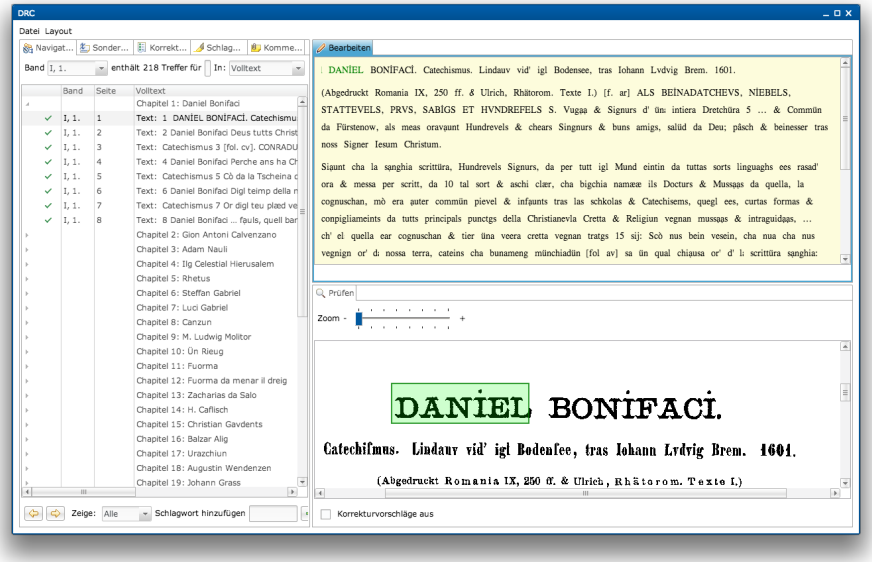

Fig. 1. The DRC editor. The pages are juxtaposed to their original digital image, with the image being linked to the text: corresponding words are highlighted in the image using the coordinates extracted after OCR.

For OCR we used *Abbyy Finereader*. By using Abbyy's pattern training facility a recognition rate of about 96,5% could be achieved, which can be rated as "average OCR accuracy" (cf. [4]). Due to the high diversity of the RC, no appropriate lexical resources were available to additionally support the OCR.

With the help of local partners in Switzerland, the project attracted about 140 users since the launch of the web page in June 2011, out of which about 20 can be considered as active participators. Until today (April 2013) more than

[4] Following a single sourcing strategy, the implementation is based on the *Remote Application Platform* (http://eclipse.org/rap/). For implementation details see [5].

[5] The digital images of the RC were provided by the project *Digizeitschriften* (http://www.digizeitschriften.de) and OCRed with *Abbyy Finereader*.

90% of the RC have been validated or corrected. During the correction phase approx. 13.7% of the running word forms were modified (about 375.000 out of 2.74 million in total).

The consequent step in building a corpus usable for research purposes is its linguistic annotation, which is subject to the follow-up project *Annotated Romansh Chrestomathy* (ARC) starting in May 2013. Analogous to the OCR correction task, we take a collaborative approach to linguistic annotation by combining automatic and manual processing steps which are conducted in a circular manner: After a first annotation cycle based on lexical lookups, the annotations are collaboratively reviewed and corrected by the language community.[6] The linguistic knowledge gained from the process of lexical annotation and manual correction serves as input to a subsequent automatic tagging cycle using the *TreeTagger* (cf. [9]). For further details on the current project see [7].

3 Collaboratively Building a Corpus - a Case Study

The overall aims of the two projects basically have two dimensions: Philologically, to provide a corpus of Romansh for research purposes, and technically, to create an integrated workbench for both OCR correction and linguistic annotation. While the technical side is described in [5] and [7], this section focuses on some methodological issues with regard to the specific language situation of Romansh. Since the correction process is still ongoing, no final evaluation can be offered, especially with regard to the linguistic annotation, which has not commenced.

A substantial result of the project is a freely accessible corpus of Romansh, that can be extended by adding further texts in a uniform manner. The philological benefit lies in the extraordinary language situation mirrored by the corpus, enabling diachronic and diatopical studies, e.g. on language contact and language change. Covering the five main idioms of Romansh and comprising different text types and genres from four centuries, the RC builds an excellent basis for a Rhaeto-Romanic text corpus (cf. [8]). Although the corpus does not take into account recent developments like the standardised language Rumantsch Grischun, its rich dialectal diversity can potentially influence the discussion about the status of the different idioms.

Besides the fact that both OCR correction and annotation can not be solved in a fully automatic manner, our main motivation for choosing a collaborative approach is the special language situation of Romansh as a minority language. Having a long and vivid tradition of language preservation and the majority of the speakers being at least bilingual, current speakers of Romansh can be regarded to have a certain expertise of their language. In combination with the text-to-image linking feature of the editor (highlighting), the expertise of being a native speaker is sufficient for this kind of proofreading task.

However, when it comes to the task of collaborative annotation this is clearly not the case. Instead of linguistic laypersons we will have to adress users with a

[6] The ARC primarily aims at part-of-speech tagging, but it is planned to subsequently add further analysis, e.g. named entities, geo-referencing, chunking and parsing, etc.

dedicated linguistic background (e.g. students, teachers) and additional tagging guidelines. Beyond this the annotation work itself will have to be coordinated to gain appropriate training data for automatic tagging. It is planned to distribute well-defined portions of text among the editors, each of these being annotated twice with the rates of inter-annotator agreement being measured (cf. [7]).

For it was one of the objectives to find out how far we can get with an completely unrestricted approach, the correction process itself was not regulated. Instead we relied on the self-regulation through the community and the editors were free to pick whatever part of the corpus they were interested in.[7] As a result, the greatest part of the RC was proofread only once, so that we cannot give meaningful measures intercorrector agreement. In retrospect, we clearly see room for improvement here, not only by coordinating the correction process more directly, but also in giving more control to the users by displaying the correction status (overall vs. per page) in the editing environment more clearly, e.g. as done in *Wikisource* or the ANDP, respectively.

4 Conclusions

This paper reports on a series of projects to collaboratively create a corpus of Romansh. Central to our approach is a web-based editor to involve members of the Romansh language community in the process of OCR correction and linguistic annotation. Although the editing environment was developed in the context of a specific use case and with regard to the specific demands arising from a multilingual, orthographically not standardised text basis, it is adaptable to virtually any other text collection. All software is published open source.[8]

The DRC project strongly supports the initial assumption that involving volunteers in the correction process is a viable option. The DRC is thus sharing the positive experiences made within the ANDP and *Wikisource*. However, some problems connected to our approach remain to be solved in future work. Although the acquisition of volunteers was successful and the participating speakers worked very thoroughly, the quality of the corrections can only be assured when the collaborative work is more clearly regulated. Especially when it comes to collaborative annotation, an unrestricted approach will certainly not be feasible. For the upcoming task of annotating the DRC it is thus essential to make the editing process more transparent to the users through explicit guidelines and by adjusting the editing environment to allow for more user interaction. Without these features, we will not achieve our goal of providing an integrated workbench for collaboratively building corpora.

[7] No personal data (e.g. cultural background, other languages) other than the optional entry of the geographical area of origin was collected. To what extent these factors might have influenced the volunteer work can therefore not be evaluated.

[8] http://github.com/spinfo/drc

References

1. Decurtins, C.: Rätoromanische Chrestomathie. Volume 1-13. Erlangen: Junge (1888-1919). Re-issued by Octopus-Verlag / Societa Retorumantscha, Chur (1982- 1986).
2. Egloff, P.; Mathieu, J.: Rätoromanische Chrestomathie - Register (Volume XV). Chur: Octopus-Verlag / Societa Retorumantscha (1986).
3. Holley, R.: Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers. National Library of Australia (2009), `http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf`
4. Holley, R.: How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. In: D-Lib Magazine 15, 3/4 (2009), `http://www.dlib.org/dlib/march09/holley/03holley.html`.
5. Neuefeind, C., Rolshoven, J., Steeg, F.: Werkzeuge und Verfahren für die Korpuserstellung durch kollaborative Volltexterschließung. In: Conference of the German Society for Computational Linguistics and Language Technology (GSCL), Hamburg (2011), p. 163-168.
6. Neuefeind, C., Steeg, F.: Stai si, defenda, Romontsch, tiu vegl lungatg - Digitalisierung als Mittel kultureller Selbstermächtigung kleinerer Sprachgemeinschaften. In: Cunha, C., Graziadei, D., Söllner, L., Stange, S., Pröbstl, T. (eds.) Über Grenzen sprechen: Mehrsprachigkeit in Europa und der Welt (language talks 2). Königshausen & Neumann, Würzburg (2012).
7. Neuefeind, C.: The Digital Romansh Chrestomathy - Towards an Annotated Corpus of Romansh. In: Zampieri, M., Diwersy, S. (eds.) Special Volume on Non-Standard Data Sources in Corpus-Based Research (ZSM Studien 5). Shaker, Aachen (2013), p. 41-58.
8. Rolshoven, J., Lutz, F., Neuefeind, C., Steeg, F.: Die Digitale Rätoromanische Chrestomathie. In: Ladinia XXXVI (2012).
9. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing, Manchester, UK (1994), p. 44-49.