# Community involvement for transcribing historical correspondences of South Tyrolean interest

## A DI-ÖSS[1] Use Case

Alexander König [1], Verena Lyding [1], Elisa Gorgaini [1], Georg Grote [2] & Monica Pretti [1]

[1] Eurac Research, Institute for Applied Linguistics
Alexander.Koenig@eurac.edu, https://orcid.org/0000-0002-8540-2396
Verena.Lyding@eurac.edu, https://orcid.org/0000-0002-2301-6860
elisa.gorgaini@gmail.com, https://orcid.org/0000-0003-3292-8914
Monica.Pretti@eurac.edu, https://orcid.org/0000-0002-3999-6103
[2] Eurac Research, Institute for Minority Rights
Georg.Grote@eurac.edu, https://orcid.org/0000-0002-3234-5623

Keywords: Crowdsourcing, Transcription, History, Linguistics, OCR, Citizen Science

**Abstract: We present a local research and citizen science initiative for the crowdsourcing-based enrichment and analysis of authentic handwritten postcards and letters, sent by South Tyrolean front-line soldiers to their families at home during the World Wars.**

In this article, we present a citizen science initiative for the documentation and transcription of historical handwritten postcards and letters by means of crowdsourcing. The documents are (mostly) authentic communications between South Tyrolean soldiers and their families at home from the First World War – during which the region of South Tyrol was right on the front lines – and the Second World War.

The presented initiative, carried out in cooperation between the Institute for Applied Linguistics (IAL)[2] and the Institute for Minority Rights (IMR)[3] at Eurac Research, is one of a number of use cases that are part of the infrastructure project DI-ÖSS[4]. The DI-ÖSS project is piloting a digital infrastructure for language data within the region of South Tyrol, which aims at exploring and exploiting synergies between actors in the language domain. This infrastructure is envisioned growing in a bottom-up manner where the special situation of the various institutions in South Tyrol that are dealing with language data is taken into account. In the currently running pilot phase, these synergistic possibilities are showcased through various specialized use cases which are implemented among the partners in the DI-ÖSS project.

[1] Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und -dienste (Digital infrastructure for the ecosystem of South Tyrolean language data and services).
[2] http://www.eurac.edu/en/research/autonomies/commul/Pages/default.aspx (23.04.2019).
[3] http://www.eurac.edu/en/research/autonomies/minrig/Pages/default.aspx (23.04.2019).
[4] http://www.eurac.edu/en/research/projects/Pages/projectdetail4262.aspx (23.04.2019).

These have been selected to represent as wide a range as possible of different types of institutions relevant to the local language situation. The partners are Eurac Research (IAL and IMR), the library "Landesbibliothek Dr. Friedrich Teßmann"[5], the language unit of the institute for culture – "Sprachstelle"[6], and the online news and community portal salto.bz[7].

This use case is related to the "letter project" which aims at collecting as many "local" correspondences as possible from the 20th century in South Tyrol, which was a tumultuous time, especially with the two World Wars and the annexation to Italy affecting the region dramatically. The collection of documents that are the basis for the project has been compiled over the past years at the IMR from local citizens all over South Tyrol. They have been actively sought out through open calls to local magazines[8], journal contributions[9] and radio broadcasts about the project. This approach is meant to put the man on the street in the focus of the historical research, investigating the lifestyle and emotions of the people by analyzing the content of the letter exchanges. So far about 13,000 items have been collected. For obvious reasons, there is a peak of material during the two World Wars because that was the time when couples and families were forced apart by circumstances which created the need to communicate via mail. This special situation also makes the collected data particularly interesting for linguistic research as it is the rare case of getting a significant amount of texts from "low literacy writers", i.e. people that probably wrote very little since they left school and suddenly found themselves in a situation where they needed to communicate via written texts. Apart from these interesting features, the special language situation in South Tyrol also means that there are many instances of code-switching to be expected within the texts, be it between standard German and the various local dialects, or – especially after the annexation – also between German and Italian. Finally, from a perspective of diachronic language change it will be relevant to compare all of these features across the course of the century.

At the current state of the project, the collaborating historian at the IMR has created a network of local families as data providers. He has been collecting several thousand missives as original documents and scanned them to JPEG-images. To enable research on characteristics of the correspondences from a linguistic as well as a historical perspective, the data needs to be made accessible not only in digital but also in machine-readable form. In particular in order to analyze the texts in a systematic way, they need to be properly equipped with metadata and transcribed. This is the part of the project that employs a citizen science approach to complete the laborious task of transcribing handwritten texts and encoding key details of the correspondences into formalized metadata. Metadata will include information about the circumstances of the correspondence, e.g. the date and place of sending, the name of the sender and the recipient, as well as their relationship (e.g. mother/son or husband/wife), and information on the quality of the writing or the script being used.

Regarding the transcription of handwritten texts with different scripts, the human contribution is of particular relevance, as software cannot easily support this. The strategy implemented in this project is to set up a crowdsourcing site to engage the local population in this project. Already during the collection of the letters, it became obvious that the people in South Tyrol are very much interested in this insight

[5] http://www.tessmann.it/en/home.html (23.04.2019).

[6] https://www.kulturinstitut.org/sprachstelle.html (23.04.2019).

[7] https://www.salto.bz (23.04.2019).

[8] Georg Grote, Ein kollektives Gedächtnis für Südtirol, in: ACADEMIA 78 (2018), 41.

[9] Mechthild Pörnbacher u.a., "Mir geht es gut ...": Feldpostkarten Südtiroler Soldaten im Ersten Weltkrieg, in: Der Schlern. Monatszeitschrift für Südtiroler Landeskunde 88/07-08 (2014), 4–21.

into the past of their region (and often also their families). Therefore, in this use case we can join the need for help from the general populace with the great interest of people in the region in interacting with their (family) history. This becomes central especially in this year, as it marks the centennial of the annexation of South Tyrol to Italy in 1919. The crowdsourcing project will be set in a context of multiple interrelated events to commemorate and highlight this historical period via engaging directly with the population and the families of those who lived through this time.

The citizen science involvement is organized in two steps: step (1) serves to systematically label the correspondences with metadata, step (2) caters for the transcription of the handwritten texts. Both steps will be completed by crowdworkers, who should be local citizens, through editing interfaces that run in standard web browsers.

For the metadata collection a local installation of the crowdsourcing framework PYBOSSA[10] has been installed at Eurac Research. PYBOSSA was chosen, because it has been released as open source[11] which meant that it could be installed by the project partners within their own IT infrastructure without having to rely on an external service provider. For the same reason it can also easily be adapted to the specific needs of this project. The interface displays the scanned document as an image and provides a number of descriptive fields related to the document (i.e. metadata such as name of the sender and the recipient, format of the correspondence, script, date, etc.), which should be completed by the crowdworkers. PYBOSSA can be configured to send the same task (i.e. providing metadata to a specific image) to multiple users and thus providing greater reliability of the results. And as the software is installed on local Eurac research infrastructure, the extraction of the results can easily be automated, and they can also automatically be converted into a suitable metadata format for further use.

The transcription of the text itself will be done in a second independent crowdsourcing phase, once the metadata has been collected. For this, several tools are currently under evaluation to find the one best suited to the needs of the project. Most likely the tool Transkribus[12] will be employed for this task. It has been developed in the context of the READ project[13] to specifically cater to the needs of archives in dealing with old handwritten documents. Transkribus is offered as a feature-rich Java application where the user can upload their documents and use built-in automatic handwriting recognition algorithms to generate a transcription, which can then further be worked on manually. While the results are often not very good out of the box, they can be used as a basis for manual transcription to provide a starting point for the crowdworkers. At the same time, those manual transcriptions can be fed back into the algorithms and improve them over time, meaning the automatic recognition will get better and better, especially in a project as described in this article where there exist a lot of texts from the same author with the same handwriting. As we want people to participate over the internet, we plan to make use of the Transkribus web interface[14], which does not provide all the functionality of the full Transkribus desktop application, but has all the features which will be needed for the transcription and has the additional advantage of providing a more streamlined interface which will work better for people more inexperienced with the use of complex transcription applications.

[10] https://pybossa.com (23.04.2019).

[11] https://github.com/Scifabric/pybossa (23.04.2019).

[12] Philip Kahle u. a., Transkribus – A service platform for transcription, recognition and retrieval of historical documents, 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto 2017, 19–24, DOI: 10.1109/ICDAR.2017.307.

[13] https://read.transkribus.eu/ (23.04.2019).

[14] https://transkribus.eu/r/read/projects/ (23.04.2019).

It has been shown in various other projects throughout Europe, for example the Dutch “Gekaapte Brieven”[15] or the large-scale European project “Europeana 1914-18”[16], that the general public is easily engaged in this kind of documentation and transcription tasks. There is quite some interest in historical letters and people enjoy interacting with the research community. On the other hand, the researchers are in need of support from the community to carry out the task of transcription. In this case, it will help to explicitly involve the local population, both because they are likely to have a stronger interest in the local history and also because they might have special knowledge (geographical or historical details) and will be familiar with the dialect variants of South Tyrolean German which can help with the transcription process. In order to correctly transcribe older German scripts (e.g. *Kurrent*), it might also be useful to specifically target an older age group that might have some experience with this kind of script. In the perspective of involving generations from the pre-digital era, Europeana has also shown that crowdsourcing does not have to stay confined to the internet but can be organized through so-called Transcribathons[17]. These are community involvement events where researchers first introduce a project through a presentation and afterwards the public can help the research process by transcribing and annotating the materials. Within the DI-ÖSS project, these kinds of events are planned in order to spark the public’s interest in the research and to encourage members of the local community to participate as equals.

## Literature

Georg Grote, Ein kollektives Gedächtnis für Südtirol, in: ACADEMIA 78 (2018), 41.

Philip Kahle u. a., Transkribus – A service platform for transcription, recognition and retrieval of historical documents, 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto 2017, 19–24, DOI: 10.1109/ICDAR.2017.307.

Mechthild Pörnbacher u.a., “Mir geht es gut ...“: Feldpostkarten Südtiroler Soldaten im Ersten Weltkrieg, in: Der Schlern. Monatszeitschrift für Südtiroler Landeskunde 88/07-08 (2014), 4–21.

http://gekaaptebrieven.nl/ (23.04.2019).

https://github.com/Scifabric/pybossa (23.04.2019).

https://pro.europeana.eu/project/europeana1914-1918 (23.04.2019).

https://pybossa.com (23.04.2019).

https://transcribathon.com/ (23.04.2019).

https://transkribus.eu/r/read/projects/ (23.04.2019).

http://www.eurac.edu/en/research/autonomies/commul/Pages/default.aspx (23.04.2019).

http://www.eurac.edu/en/research/autonomies/minrig/Pages/default.aspx (23.04.2019).

http://www.eurac.edu/en/research/projects/Pages/projectdetail4262.aspx (23.04.2019).

https://www.kulturinstitut.org/sprachstelle.html (23.04.2019).

https://www.salto.bz (23.04.2019).

http://www.tessmann.it/en/home.html (23.04.2019).

[15] http://gekaaptebrieven.nl/ (23.04.2019).

[16] https://pro.europeana.eu/project/europeana1914-1918 (23.04.2019).

[17] https://transcribathon.com/ (23.04.2019).