Handwritten Text Recognition - *Transkribus*: A User Report

The electronic Text Reuse Acquisition Project (eTRAP)

Melina Jander
melina.jander@stud.uni-goettingen.de

November 2016

Since it took off in July 2016, the pilot project **Tracing Authorship in Noise (TrAIN)**[1] has made notable progress in the process of Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR). For the latter process, we are using the current state-of-the-art tool *Transkribus*, a comprehensive transcription and recognition platform currently and mainly used by humanities scholars.[2]

Transcribing historical handwritten texts is not only a highly interesting but also a challenging task. Transcriptions need to be carried out very thoroughly – otherwise *Transkribus* might not 'learn' as much as it should. But how does the programme 'learn', actually? To put it simply, *Transkribus* 'learns' to 'read' handwritten texts of a certain author by 'seeing' as much of his or her handwriting as possible. Take the letters of Jacob Grimm as an example[3]: we selected 52 letters Jacob Grimm wrote to his brother Wilhelm and to some of his acquaintances; the digitised images of these letters were uploaded to the platform so that our research assistants (see Project Information) could begin working. They did not only transcribe the handwritten words but they also added relevant information about both the style and the content of the letters. To do so, *Transkribus* provides its users with valuable tools, which are briefly explained in reference to our TrAIN project.

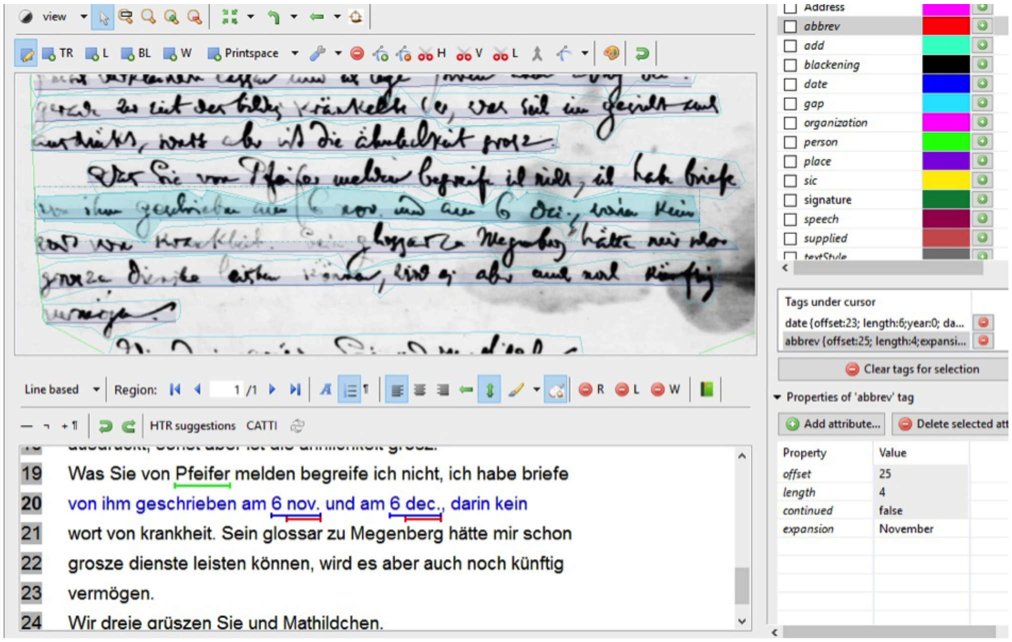
Firstly, we define the parts of the letter to be transcribed: the text regions, its lines and their baselines. This information gives us a frame for the transcription and forces us to understand the reading order of the letter. When reading a letter, we are often faced with interline additions. We humans intuitively integrate those additions into our reading flow, but a programme like *Transkribus* lacks such an intuition - it needs help. What a transcriber can do to keep the reading flow linear is to integrate the interline additions by inserting extra lines and baselines for them. In this way, the transcriber

[1] For more information about the project, see: http://www.etrap.eu/research/tracing-authorship-in-noise-train/ (Accessed: 15 October 2016).

[2] Further information is available at: `https://transkribus.eu/Transkribus/` (Accessed: 18 October 2016).

[3] The sample data for the TrAIN project contains 85 letters of Jacob and Wilhelm Grimm. The letters pertain to both their adulthood –Jacob's are much easier to read than Wilhelm's– and childhood.

becomes a creator because she or he can split the original lines, add the new line (for the interline addition) and correct the numbering of the lines, thus enabling *Transkribus* to know and show its users the correct reading order. After these steps, the transcriber can finally begin with the writing process. Since *Transkribus* is designed to keep the transcriptions as tidy as possible, every detected line of the handwritten letter is tied to its equivalent in the text editor – getting lost or producing an untidy transcription is nearly impossible. You might wonder what a person like Jacob Grimm –an intellectual with a passion for the German language– was writing about in his letters. He discussed both professional and private topics, mentioning different people, various places and, of course, a wide range of dates. Such information is important to get an understanding of Jacob Grimm's stylistic evolution, which is strongly linked to his work and personal experiences.



19 Was Sie von Pfeifer melden begreife ich nicht, ich habe briefe
20 von ihm geschrieben am 6 nov. und am 6 dec., darin kein
21 wort von krankheit. Sein glossar zu Megenberg hätte mir schon
22 grosze dienste leisten können, wird es aber auch noch künftig
23 vermögen.
24 Wir dreie grüszen Sie und Mathildchen.

□	Address	🟪
□	*abbrev*	🟥
□	*add*	🟩
□	*blackening*	⬛
□	*date*	🟦
□	*gap*	🟦
□	*organization*	🟪
□	*person*	🟩
□	*place*	🟪
□	*sic*	🟨
□	signature	🟩
□	*speech*	🟥
□	*supplied*	🟥
□	textStyle	⬛

Tags under cursor
date {offset:23; length:6;year:0; da...
abbrev {offset:25; length:4;expansi...

Properties of 'abbrev' tag

Property	Value
offset	25
length	4
continued	false
expansion	November

Figure 1: Screen-shot of a transcription in *Transkribus*. Line 20 shows that an entity can entail more than one tag (here: tags for date and abbreviation).

The Tagging function allows the transcriber to add details to certain entities, in our case mainly people, places, dates and abbreviations. By using this tool, the transcriber feeds the programme with more information, which, at some point, leads it to recognise particular words, e.g. the abbreviation 'u.' for the German word 'und'[4]. While the Tagging tool adds information about the content to the transcription, the Metadata function is very useful for further stylistic information, for it enables the transcriber to keep her or his transcription as close to the text image as possible. If, for example, Jacob Grimm struck through some words in his letters, we can mark those words as Strikethrough; thus, Transkribus learns that not every entity it detects as a word belongs to the content of the letter, even though it clearly belongs to the text and, therefore, to the transcription.

Teaching *Transkribus* through manual transcription helps it run automatic Handwritten Text Recognition, which is a great acceleration of our work-flow. The programme enjoys constant developments, which make it an easier and faster experience;[5] that is why –speaking from our point of view– *Transkribus* proves as a very useful tool for HTR, especially because it is easy to handle but, nevertheless, advanced enough to serve the goal of our TrAIN project.

To cite this report, and depending on your preferred referencing style, you can adapt the following:

Jander, M. (2016) *Handwritten Text Recognition – Transkribus: A User Report*, eTRAP Research Group, Institute of Computer Science, University of Göttingen, Germany, 2 November. Available at: http://www.etrap.eu/academic-output/.

Project Information

Name: Tracing Authorship In Noise (TrAIN)
Number: 392860 "Campuslabor"
Grant: 20,000 €
Funder: University of Göttingen, Campuslabor-Digitalisierung
Duration: 1 July 2016 - 31 December 2016
Contact: contact@etrap.eu

Research Assistants: Melina Jander, Svenja Walkenhorst, Linda Brandt
Supervision: Gabriela Rotari, Emily Franzini

[4] According to the *Transkribus* Wiki, at least 2000 lines are needed to start the automatic HTR. But: The more text is transcribed, the better the HTR results of *Transkribus* (https://transkribus.eu/wiki/index.php/Main_Page, accessed: 18 October 2016).

[5] Furthermore, the *How To Papers* are a very helpful user's guide comprising simple explanations and descriptions, as well as screen-shots. You can download these papers from https://transkribus.eu/Transkribus/).