Optical Recognition Assisted Transcription with Transkribus: The Experiment concerning Eugène Wilhelm's Personal Diary (1885-1951)

Régis Schlagdenhauffen

Journal of Data Mining and Digital Humanities

Optical Recognition Assisted Transcription with Transkribus: The Experiment concerning Eugène Wilhelm's Personal Diary (1885-1951)

Régis Schlagdenhauffen[1]

1 EHESS, Paris, France

*Corresponding author: regis.schlagdenhauffen@ehess.fr

Abstract

This article proposes use the Transkribus software to report on a "user experiment" in a French-speaking context. It is based on the semi-automated transcription project using the diary of the jurist Eugène Wilhelm (1866-1951). This diary presents two main challenges. The first is related to the time covered by the writing process - 66 years. This leads to variations in the form of the writing, which becomes increasingly "unreadable" with time. The second challenge is related to the concomitant use of two alphabets: Roman for everyday text and Greek for private issues.

After presenting the project and the specificities related to the use of the tool, the experiment presented in this contribution is structured around two aspects. Firstly, I will summarise the main obstacles encountered and the solutions provided to overcome them. Secondly, I will come back to the collaborative transcription experiment carried out with students in the classroom, presenting the difficulties observed and the solutions found to overcome them. In conclusion, I will propose an assessment of the use of this Human Text Recognition software in a French-speaking context and in a teaching situation.

Keywords:

Human Text Recognition – Learning process – OCR – TEI – User Experiment

INTRODUCTION

This article offers feedback and thought about a collaborative transcription experiment conducted at the Ecole des Hautes Etudes en Sciences Sociales (EHESS) since 2017, notably in the framework of a seminar conducted with Master students[1]. It is part of the field of research in digital humanities [Mounier, 2018], is based on the concept of user experience (UX) and questions the use of tools for digital humanities. These questions have been studied among others by Gibbs and Owen who shows that until the 2010's, "despite significant investment in digital humanities tool development, most tools have remained a fringe element in humanities scholarship" [Gibbs, Owen, 2012] and by Warwick who considers that, in this field, "we must make sure that users know that they do not need to do their job. If the digital resources correspond well to what they want to do with them, the users adopt them" [Warwick, 2012]. More recently, Muehlberger *et al.* [Muehlberger *et al.*, 2018] have published the very first overview of research on how to transform scholarship in the archives through handwritten text recognition. In their contribution, they reported on some outputs of the European Union's

[1] https://enseignements-2019.ehess.fr/2019/ue/2525/

Horizon 2020 project on the recognition and enrichment of archival documents (READ), which develops advanced text recognition technology based on artificial neural networks: the Transkribus platform.

In a French-speaking context, [Massot *et al*., 2018] were the first to propose a critical study of the use of this application in the setting of the assisted transcription of Michel Foucault's reading sheets. Following this, [Perrin, 2019] proposed a tutorial on the collaborative transcription of the archaeological archives of the Bibracte site which restores the user experience applied to this tool, point by point[2].

The "user experience" designates all perceptions, interactions and feelings that the user experiences with respect to a product or service before, during and after its use [Christine, Trognot, 2015]. When applied to handwriting recognition software, it can refer either to printed material or to hand-written manuscripts. The first case involves OCR (*Optical Character Recognition*) software whose use is now developed in many fields. The second involves HTR (*Human Text Recognition*) software, a field which is still in full development as it presents additional challenges compared to OCR: the singularity of each human handwriting style vs. font standardization, variability of handwriting forms, use of abbreviations, presence of erasures, blacking out, etc. Furthermore, while OCR software is already relatively well-established [Schantz, 1982], software designed to automate all or part of the transcription process of human handwriting remain rare. One such software, Transkribus, was developed by a consortium of academics within the framework of two successive European H2020 programmes (Recognition and Enrichment of Archival Documents, READ project)[3]. By means of an expert tool (*expert client*), the software developed since 2016 allows the downloading of documents and images within a private but shareable collection. Images are segmented into blocks, then lines, and finally words using layout analysis tools. Text can be linked to images then transcribed into any language with the character font chosen. In the end, the software allows exportation of transcribed documents in several formats (DOCX, TEI, RTF, PDF, XML).

The rest of the paper will be organized as follows. Firstly, we will present the TransDiary-TEI project[4] and its specificities with respect to the Transkribus application. We will address issues related to the distinction of the corpus used and to the continuous improvement of the use of the tool. Then, we will return to the user experiment by discussing the difficulties encountered. Following this, we will give an assessment of the use of the Transkribus software applied to the particular scope of the TransDiary-TEI project.

1 THE TRANSDIARY-TEI PROJECT

1.1 The characteristics of Eugène Wilhelm's diary

The personal diary of the jurist Eugène Wilhelm (Strasbourg, 1866-1951) is in the form of 55 numbered notebooks, mostly covered with moleskin. These total 8538 pages written in French, but illuminated with fragments of the Greek alphabet. The diary was written over 66 years (from age 19 to 85 of the diarist) with great regularity. It provides contextual elements of his professional career, his daily life, his intellectual trajectory and the European political situation between 1885 and 1951. It also provides an account of the diarist's sexual relationships with

[2] https://f.hypotheses.org/wp-content/blogs.dir/7177/files/2019/11/Tutoriel_Transkribus_V2.pdf
[3] https://read.transkribus.eu
[4] https://cahier.hypotheses.org/transdiary-tei

men and women of all ages and social conditions. The notebooks vary in size and cover five major historical periods: the Kaiserreich - the period during which Alsace was governed as part of the German Empire, i.e. notebooks 1 to 25 (3146 pp.), the period corresponding to the First World War, notebooks 26 to 30 (906 pp.), the inter-war period, notebooks 31 to 41 (2200 pp.), the Second World War, notebooks 42 to 51 (1416 pp.) and the post-war period, notebooks 52 to 55 (870 pp.). The specific features of Eugène Wilhelm's diary were presented by [Dubout, 2016 and 2018], [Dubout and Schlagdenhauffen, 2014] and [Schlagdenhauffen, 2014 and 2015]. They are comprised of summaries written subsequently by the diarist, at the beginning of each notebook, annual reports and are written in two distinct alphabets. The Roman alphabet represented ¾ of the diary, and the Greek alphabet ¼, whilst the whole diary was written in French.

To make this extraordinarily rich diary accessible to as many people as possible, the TransDiary-TEI project had the idea of offering a digital edition. Such a project has various advantages, especially because of the size of the diary, the length of time over which it was kept, and the possibility of continuously improving transcription quality while enriching it with metadata. The last two components of the project concern named entities (proper names of persons and places[5]), which may lead to confusion, homonyms or the necessity of additional information as the research evolves. Also, metadata relating to cross-references inside and outside the corpus was concerned, allowing monitoring of the diarist's path through time and space, mapping of his networks, and linking with international databases or specialized sites to create a global and digital universe around Eugène Wilhelm. Finally, a digital edition will provide access to the materiality of the documents stored in the EHESS research data warehouse, *Didomena*[6]. In the context of the TransDiary-TEI project, as an EHESS team we transcribe the diary using the Transkribus tool, and it is this experiment that will be reported in the following sections.

1.2 Specificities of the Transkribus tool applied to the project

Personal diaries are a field that is still barely explored either by human handwriting recognition (HTR) software or by TEI unlike letters, registers, or public archive documents [Cummings, 2008]. Also, in the case of Eugène Wilhelm's diary, two phenomena must be taken into account simultaneously: the duration of the writing and its variability. Indeed, between the diarist's 19th and 85th birthdays, writing evolved in a way that makes it difficult to establish a single training model for the moment. Indeed, writing was regular and, so to speak, "scholastic" at the beginning of the period, but it became more complex with age, and was difficult to read by the end of the diarist's life (Figure 1).

[5] *Cf.* for example [Brando *et al.*, 2018] "Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle". https://hal.archives-ouvertes.fr/hal-01925816/file/8.Article-SAGeo_adaptation-et-évaluation-de-systèmes-de-REN-et-NEL.pdf

[6] https://didomena.ehess.fr

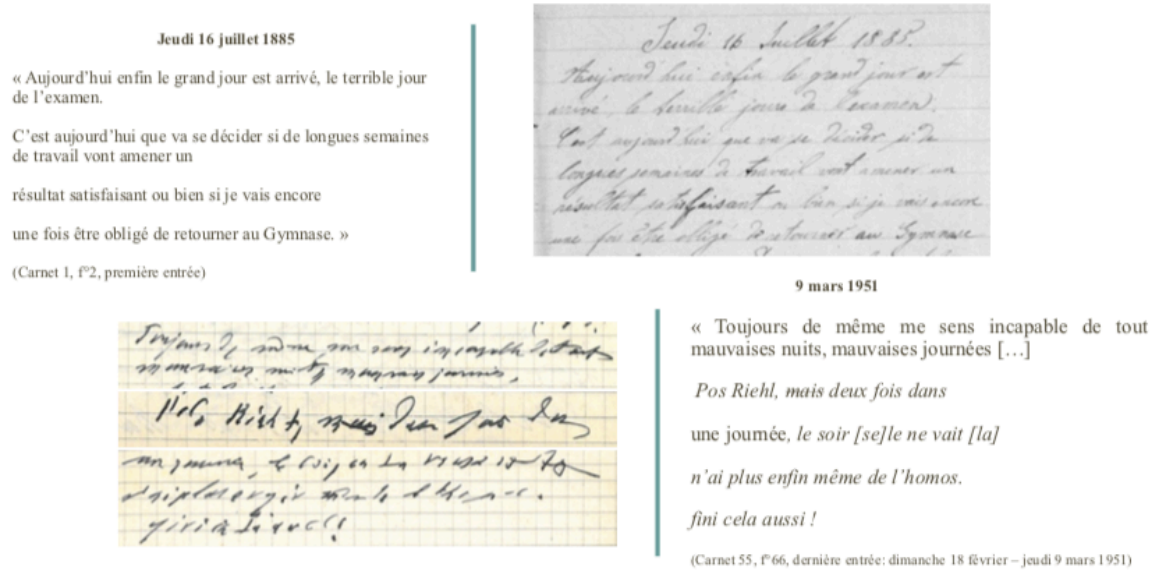

Figure 1. *Variations in Eugène Wilhelm's writing, a comparison of notebook 1 (1885) and notebook 55 (1951).*

The quasi-alternating use of the Latin and Greek alphabets also has to be taken into account (Figure 2).

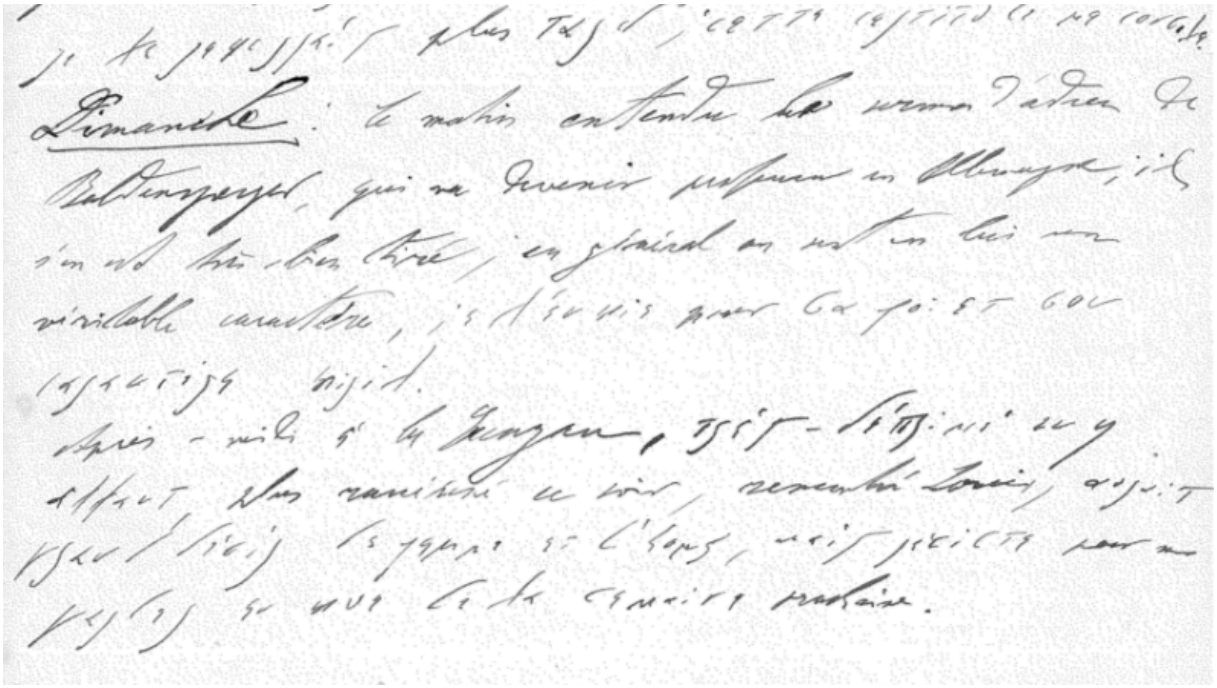

Figure 2. *Alternating the Greek and Latin alphabets, notebook 10, 1890, f°6/68.*

Bearing this in mind, the first tests carried out in 2017 indicated a high error rate (40% in training mode and 49% in validation mode), which we have since been working to reduce. Especially, we used a suitable dictionary, but we have also been training more efficiently and have paid attention to writing variation over the long-term.

Concerning another aspect of the project, some models already exist that relate to encoding in TEI. However, these remain unsatisfactory as pointed out by [Nelson, 2017] or difficult in French as reiterated by others [Soudan, 2012]. Nevertheless, apart from these considerations the first step was to produce transcripts to be able to train a model, as recommended in the Transkribus user guide[7]. This is the case for other automated transcription projects using HTR software.

1.3 HTR - use and pattern training

To train a model for automated transcription, we first delivered about 100 pages to the READ-Transkribus consortium team. At this stage, the results were not conclusive since the error rate was higher than 40% (Figure 3).

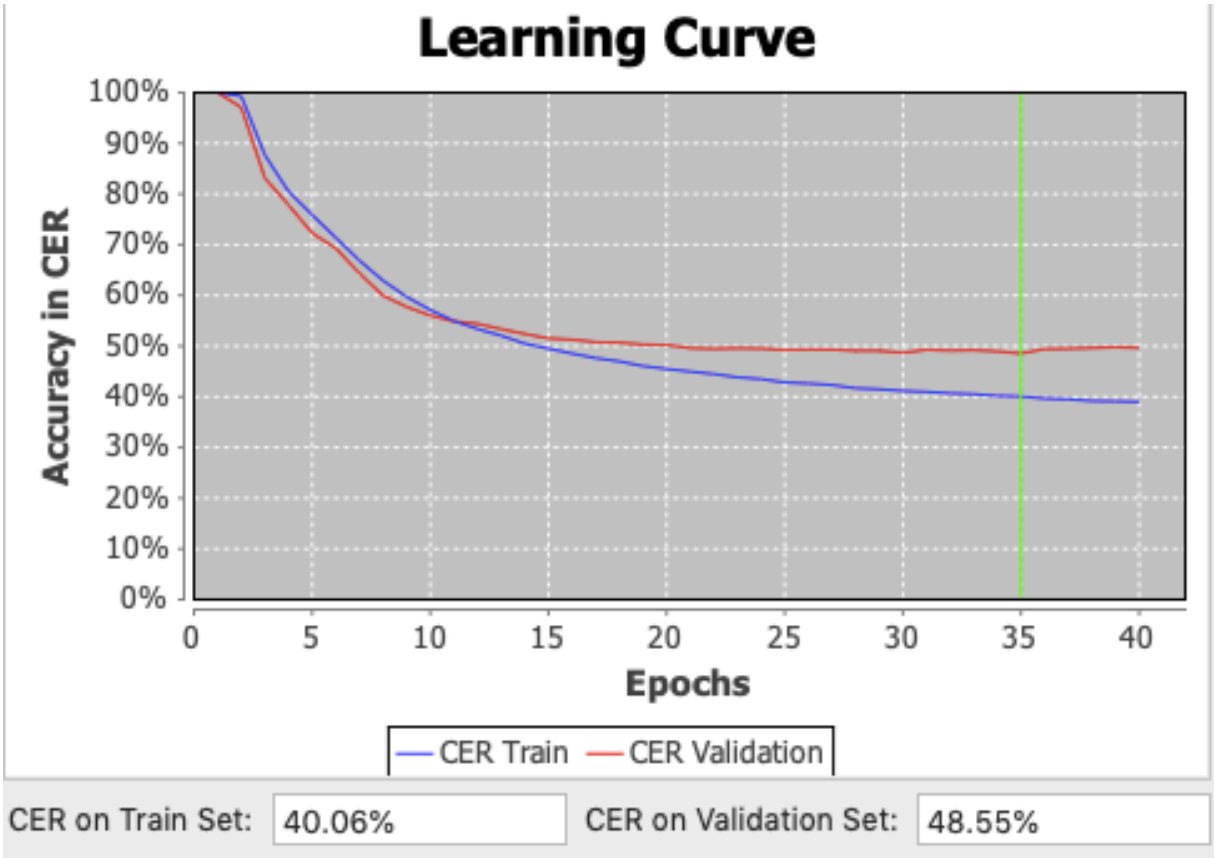

Figure 3. *Model of "Eugene Wilhelm t2i_M1", realized on 02.02.2018.*

The first model developped by the Transkribus team, was known as the "Eugene Wilhelm t2i_M1 model" (Figure 3) and automated recognition remained laborious using it. Furthermore, the error rate stayed unsatisfactorily high. However, from the creation of this first model, we observed that HTR recognized Greek characters more easily than Roman characters. One reason for this is that letters written in Greek are detached rather than linked to one another.

Another explanation lies in scanning quality. Low-resolution and mostly black and white scans made automatic line recognition difficult. At this stage, automatic baseline recognition was abandoned in favour of manual tracing of text boxes and lines, which remained an efficient but

[7] https://transkribus.eu/wiki/images/7/77/How_to_use_TRANSKRIBUS_-_10_steps.pdf

time-consuming solution. The first model was not backed up by a dictionary and showed the limits of automatic recognition of human handwriting in French. Indeed, as pointed out by [Massot *et al*., 2018]: "Transkribus does not use a dictionary and does not attempt to recognize words as it analyzes lines of text *character by character*." Thus, the first automatic transcriptions made without a dictionary transcribed a little-known language, worthy of Voynich's manuscript!

As a consequence, the expected effects of automatized transcription did not make themselves felt during this first phase. The involvement of Carmen Brando, PhD in computer science and head of the EHESS geomatics platform, together with the advice of Joachim Dornbusch from the EHESS Digital Research Centre was a decisive moment. We agreed to create a Eugène Wilhelm dictionary to facilitate automated transcription based on earlier transcriptions. We presumed that the use of a specific dictionary presupposed that the vocabulary of the transcribed part was more or less homogeneous with that of the part to be transcribed. This choice was recommended by the Transkribus team, who considered that "*there is one thing to take into account when using a (general) dictionary: The bigger (in the number of words) the dictionary is, the slower the recognition process will be. Therefore, a general dictionary is often not really usable*."

The creation of the dictionary consisted of morphosyntactic tagging by means of TXM processing (lexicon-words.txt). This enabled us to create a list of words (lexicon-occurrences.csv). The process was carried out using the vocabulary used by Eugène Wilhelm, starting from the texts that had already been transcribed (about 2000 pages out of more than 8000). To do this, three consecutive processes had to be carried out. Firstly, we extracted a list of words from the texts by a process of segmentation and tokenization of the texts; secondly, we proceeded with a process of lemmatization, so that each word was associated with its grammatical category; thirdly, we flexed each word (except for empty words and proper nouns) beginning with the rules of French grammar.

In the end, the dictionary had more than 34,000 flexed forms and was a powerful resource.

By choosing a dictionary that was specific to this corpus, we had correctly guessed that the tool would gain in efficiency and speed.

In fact, from this point we were able to create a new, more efficient, automated transcription model called the "EW_2+" model. This model has an error rate of 5.27% in the training mode and 18% in the validation mode (Figure 4). The latter model is still used, and has led to significant improvements in transcription, demonstrating the efficiency of the tool and paving the way for a satisfactory collaborative transcription experience.

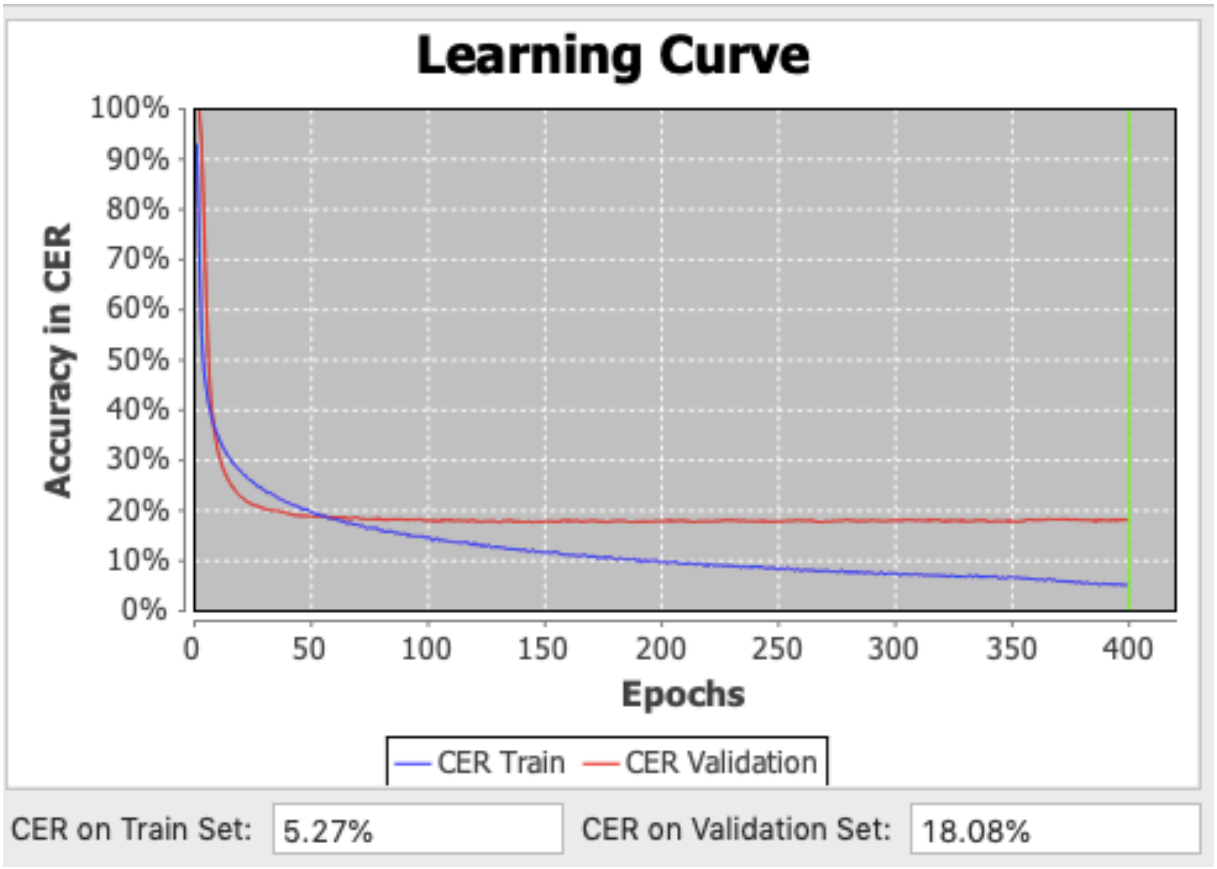

Figure 4: EW_2+ model, realized on 15.12.2018.

2 COLLABORATIVE TRANSCRIPTION WITH THE STUDENTS

In this section, we would like to return to the collaborative transcription experiment conducted with EHESS Master students. At the beginning of the seminar, these students did not necessarily have any prior knowledge concerning the use of transcription software. The 24-hour/semester course, i.e. 2 hours per week, has now been run for 3 years at the EHESS. Its objective is two-fold: firstly, to initiate students in the use of the tool, and secondly, to enable them to master all the steps up to the training of a model. In this context, we used Eugène Wilhelm's diary as a support for the different steps.

2.1 Teaching sequences

To facilitate the exercise, we started with one of the first notebooks of the corpus, notebook n°3 (from August 7th 1886 to June 5th 1887), Like the first other 10, notebook n°3 has the specificity of being “readable” as far as the parts written using the Roman alphabet are concerned. The parts written in Greek remained difficult for the students because of the absence of compulsory teaching of Ancient Greek in French schools. Accordingly, teaching is divided into the following pedagogical sequences:

The first pedagogical sequence is a general presentation of the interface. The students were invited to take part in a "collection" specially created within the framework of the seminar. This grouped together several notebooks dedicated to manual training. The whole logbook had already been downloaded onto the application within the framework of the TransDiary-TEI project. This phase consists of a first familiarization with the tool and the corpus.

The second stage of teaching is dedicated to the manual initiation to tracing text zones and then baselines. This approach makes it possible to learn these operations, which can later be automated but still require adjustments if necessary. The ability to trace text boxes and lines manually is of great importance, as it allows compensation for possible shortcomings of the tool.

The third step consists of automating the two processes described above and making the necessary adjustments (retracing certain lines, adding or deleting lines that are not recognized or are wrongly recognized by the tool). This operation can be made easier by displaying in Transkribus the line numbering and/or other options that allow an optimal adjustment of the interface according to the expectations of the moment.

The fourth step consists of increased habituation to diaristic writing. In this context, an initiation to, or a reminder of, the Greek alphabet helps students who wish to understand everything. At the same time, the pages of the diary are read collectively to recognize the writing and to discuss the best interpretation of confusing words. Here, this only involves proper names of places and people. This step may seem the simplest because it does not require any technical knowledge but in fact it is very complex because it requires habituation. It was spread out over several sessions.

A fifth step was to use an existing HTR model to experiment with the power of the Transkribus tool. This step provided access to the tool's collaborative dimension. It is undoubtedly one of the most stimulating steps. During each session, a student was in charge of directing the operations carried out on the page being studied. Indeed, only one person at a time can make modifications and save them. This is perhaps one of the current limitations of the tool when several people are working collectively on the same page. Once the automated transcription was completed, the operations were threefold. Firstly, the lines had to be adjusted (certain lines were lengthened or shortened and/or lines that were not correctly recognised were added or deleted). Secondly, misrecognized words had to be corrected. Such words could fall into any category: incorrectly conjugated verbs, mismatched adjectives, unrecognized words. During this phase of recognition and interpretation we collectively discuss the best interpretation of a word that is difficult to read. When there is still a source of doubt and hesitation, we add an "unclear" tag. The question of tagging leads us to consider the third operation, which was related to the enrichment of the text by metadata. By default, Transkribus proposes a certain number of tags such as: abbreviation, address, date, place, organization, person, etc.

The sixth step of the seminar is the use of transcript export functions. This is particularly interesting because it allows the selection both of different formats (docx, pdf, tei) and of different options such as ALTO/METS, BIO, TEI standards. Alternatively, exports can be used to keep the line and page breaks, to mark words as "unclear" and more generally to keep selected tags. All these options are intended to facilitate the digital editing of a source (Figure 5).

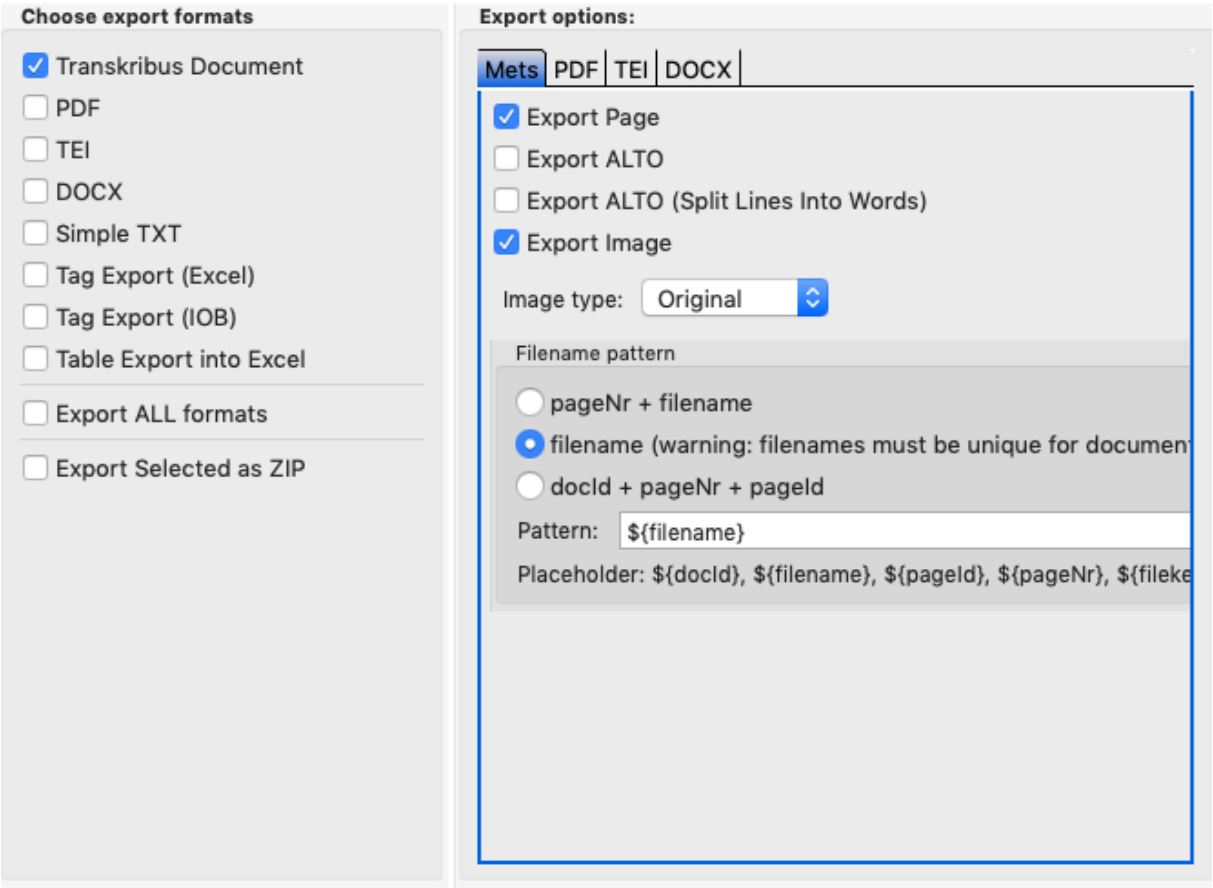

Figure 5. Export formats.

Finally, a seventh step consists of the pattern training of a model (Figure 6)[8]. For the time being, only training of models derived from models previously created have been carried out within the framework of the course. They can be trained using the collection of available basic models to improve training. This seventh step is intended to introduce students to modelling training but it has not yet been carried out with autonomous corpora.

[8] For more information on model training, see: https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf

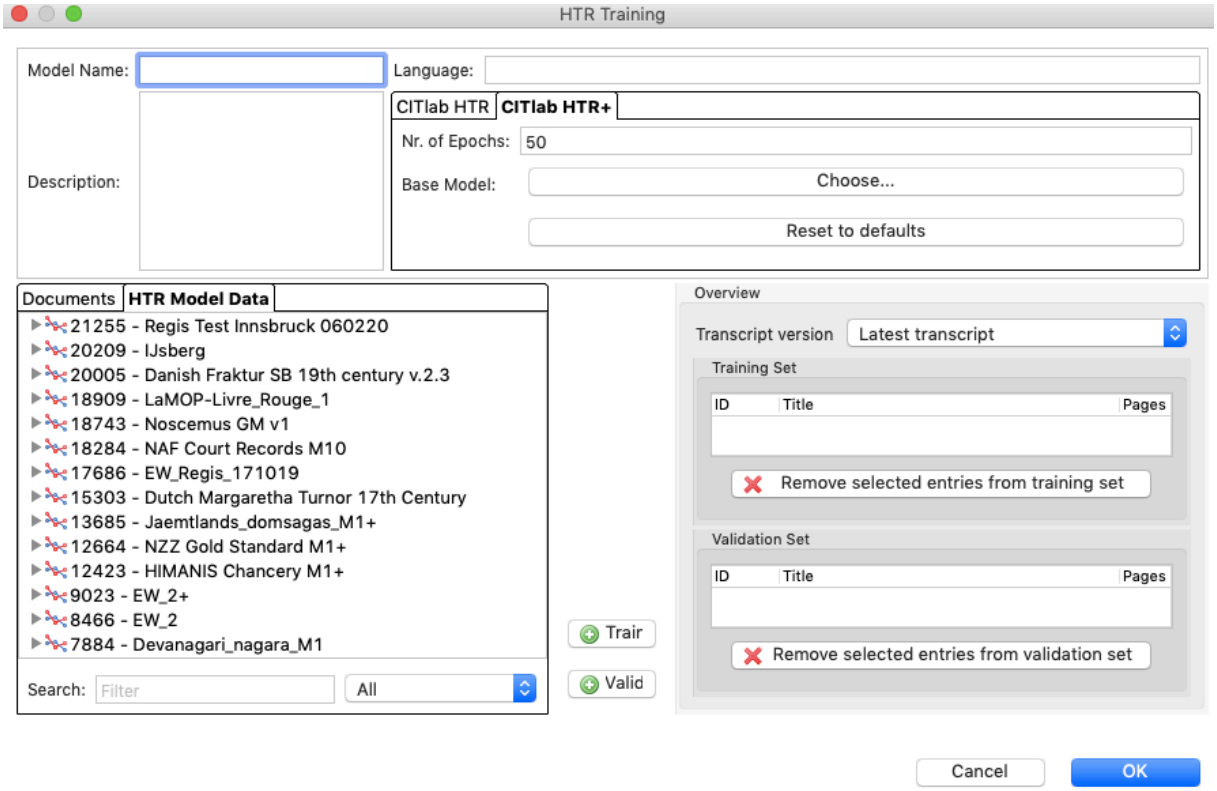

Figure 6. The interface of model training.

After this presentation of different instructive sequences of an initiation seminar on the Transkribus tool's use, in the following section we will examine the difficulties encountered and the solutions deployed to solve them.

2.2 Difficulties encountered and solutions deployed

Several difficulties have been encountered so far. The first of these is linked to the general readability of the diarist's writing. In the project to transcribe Michel Foucault's notes [Massot et al., 2018], the diarist was found to truncate or shortens certain words or does not apply himself in keeping his diary. In the context of teaching methodology applied to the Transkribus software, this presents a double challenge. It is necessary either to work on parts of the diary that are immediately readable (the so-called youth diary), or to work on diary entries that are more difficult for the human eye to decipher. In the former case, the tool's power is not particularly noticeable so this can be substantiated. However, this makes verification by the human eye more complex, especially as the eye is not accustomed to the specificity of the diarist's writing. Another singularity is added, namely the use of the previously mentioned Greek alphabet. It was agreed that passages in Greek should be transcribed into the Roman alphabet because of the current limits of understanding of the Greek alphabet.

Another difficulty is linked to the students' lack of knowledge concerning the general political and historical context of the diarist. For example, the Reichsland period for Alsace (1871-1914) that forms the backdrop for notebook n°3 requires contextualization. This problem can also be considered as an advantage since it allows for exciting class discussions and the acquisition by students of new knowledge on a subject that is little-studied, at least in the French context. In retrospect, it seems that discussions are among the most exciting moments. These may be events

recounted by the diarist we are commenting on, or the possibility of following him step by step thanks to the cartography of the places he has visited or passed through. We pursued his movements, restoring their logic. Within the tool, the difficulty linked to the monolingualism of the interface (in English) does not seem to have played a determining role. Once the various functionalities were understood and learned, it was possible for the students to test them, reproduce them, and ultimately understand the general logic that prevails at this level of the application's use.

In addition, the limitations encountered by the absence of a dictionary integrated into Transkribus should be considered. As we pointed out above, the first automated transcriptions of Eugene Wilhelm's diary were of poor quality because there was no dictionary. To correct this, we opted for the creation of a dictionary of the words used by Eugène Wilhelm (see section 1.3)[9] rather than using a dictionary of the French language. Indeed, as each individual's vocabulary remains limited - or circumscribed - to certain spheres, we saw no point in backing up transcriptions with a general dictionary. For example, vocabularies from the fields of natural sciences, medicine or crafts, are all unused, and therefore unnecessary, words in the context of writing a lawyer's diary.

Finally, there was the question of the cost/benefit ratio according to the size of the corpus. In our case, part of the corpus had been the subject of a first transcription carried out mainly by Régis Schlagdenhauffen and Kevin Dubout[10]. Indeed, about 2000 pages or ¼ of the approximately 8000 pages corpus had been previously transcribed. Thus, importing the learning data into the application was easy. As a first step, a test was carried out by the Transkribus team, with 200 images drawn from several of the diary's notebooks. The prevalent idea at that time was to immediately reinstate the writing variations that occurred in the notebooks. The use this strategy made it possible to accelerate considerably the production of a first *training dataset*. Nevertheless, in the case of students who often work on smaller corpora as part of their dissertations, the question of the time and energy required remained unresolved. Paradoxically, the use of the OCR tool probably seems to be the most accessible for students in the immediate future, since they also work on printed corpora (Figure 7). The OCR tool is indeed extremely easy to use.

[9] The author thanks Carmen Brando (EHESS) for creating the Eugène Wilhelm dictionary.

[10] The author is particularly grateful to Nicolas Eybalin and Sara Maïka for their help during the first phase of transcription, and to Günter Hackl, Günter Mühlberger, Marie-Laurence Bonhomme and Carmen Brando for the training phases.

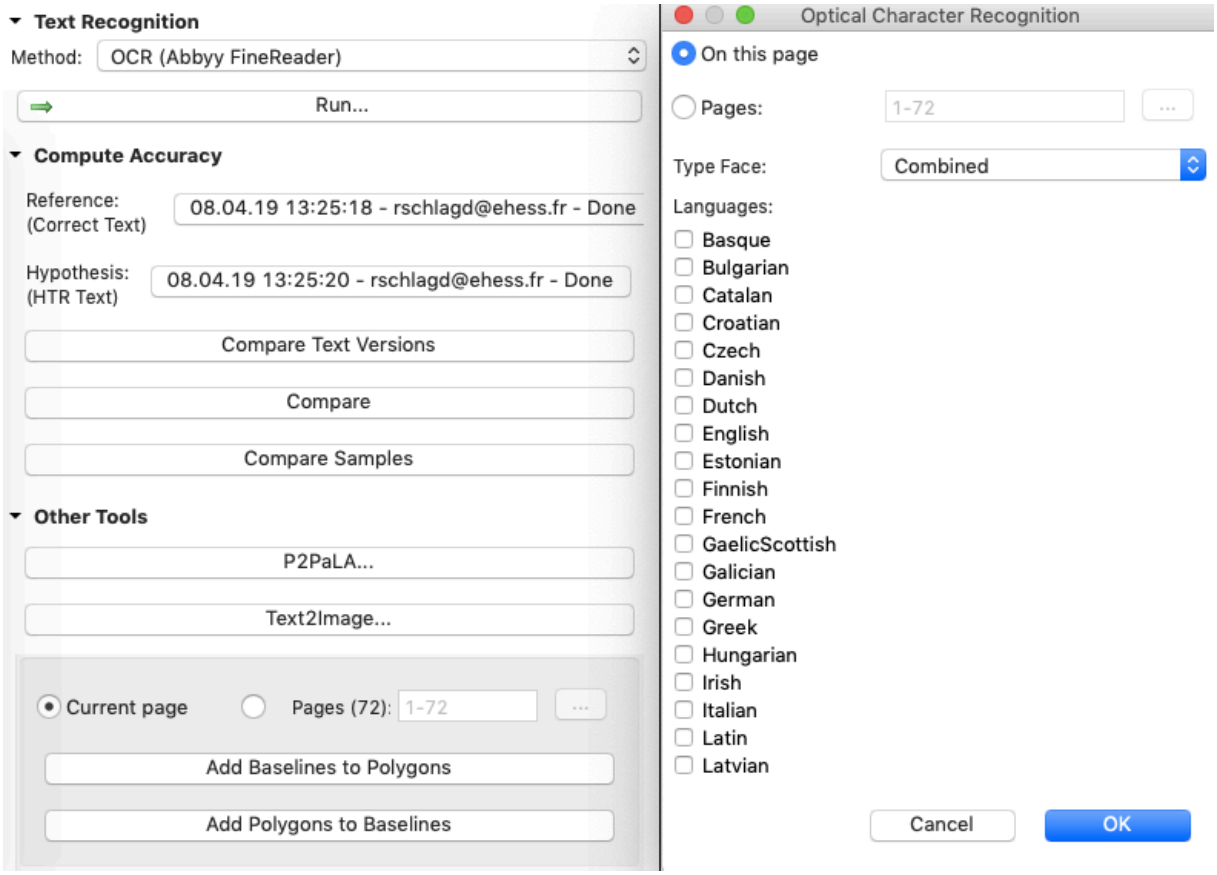

Figure 7. Screenshot of the OCR module.

Compared to other tools, the Transkribus OCR feature has the advantage of being able to back up scans and transcripts to a collection stored on the server. This is particularly useful, for example when working on a periodical, or on various printed publications of one author. In addition, the OCR tool offers export functionalities (format variability, support for tags, tags, etc.). Moreover, because OCRisation is now a well-established feature, recognition rates are high - even without the use of pre-existing training data. Finally, this tool, like the HTR tool, can easily be reused, for example for semi-automated translation. Also, it is perhaps the possibility of being able to transcribe both handwritten and printed texts that gives Transkribus its remarkable power.

CONCLUSION

By way of assessment, we can only confirm some of the limitations observed by [Massot et al., 2018]: "automated transcription must be post-edited manually by proofreaders, but the results obtained remain positive because automatic transcription allows faster manual transcription and helps in the recognition of certain words". Moreover, despite their imperfections, automatic transcriptions are already usable for "full-text" searches, which is considered in many ways to be one of the most relevant features of Transkribus. Finally, "Transkribus does not use a dictionary for the automatic transcription phase and instead analyses manuscripts letter by letter. Thus, the results could be improved by using automatic correction algorithms that search for similarities to 'clean' the automatically produced data" [Massot *et al.*, 2018]. It should

therefore be remembered that only high-definition scans and the use of a model with a low error rate (less than or equal to approximately 5%) make for a pleasant user experience. In our case, the difficulty related to the reliability of recognition by the model has been partially circumvented due to the use of a suitable dictionary. However, at present this remains almost insurmountable for any other corpus in French.

In terms of user experience, as long as the interface was simple we did not find any particular difficulty for the students. In fact, the singularity of the diarist's writing seemed to present difficulties for the students more than it was not so much the tool itself. Finally, one of the persistent limitations for its reuse on an individual basis in the framework of smaller projects lies in the existence of the few reusable public models requiring the realization of a new model[11]. However, as we pointed out, the use of the OCRisation tool remains immediately accessible and it is undoubtedly at this level that students can find immediate satisfaction. Consequently, the balance is mixed regarding the user experience applied to students in humanities and social sciences (i.e. non-archivists). Indeed, Transkribus is only effective for medium or large manuscript corpora. Its use is probably too expensive for anyone trying to carry it out on a small corpus or one with a wide variety of writers. However, the more the software is used and the more templates are made available, the easier semi-automated recognition of a wide variety of handwriting will become.

[11] The model developed as part of this work is not yet open because the error rate is still too high. We hope to make it public soon, following further HD digitizations that will allow us to re-train the model.

REFERENCES

Brando C., Soudani A., Meherzi Y., Bouhafs A., Frontini F., Dupont Y. and Melanie-Becquet F. Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle. *Atelier Humanités Numériques Spatialisées (HumaNS'2018)*. 2018. https://hal.archives-ouvertes.fr/hal-01925816

Brent N. Curating Object-Oriented Collections Using the TEI. *Journal of the Text Encoding Initiative*. 2017;9. https://journals.openedition.org/jtei/1680

Cummings W. The William Godwin's Diaries Project. *Jahrbuch für Computerphilologie*. 2008;10. http://computerphilologie.de/jg08/cummings.pdf

Dubout K. Durch Rezensionen zur Emanzipation? Die „Bibliographie der Homosexualität" (1900-1922) im Jahrbuch für sexuelle Zwischenstufen. *LIBREAS. Library Ideas*. 2016;29. https://edoc.hu-berlin.de/handle/18452/9746?show=full

Dubout K. *Der Richter und sein Tagebuch. Eugen Wilhelm als Elsässer und homosexueller Aktivist im deutschen Kaiserreich*, Campus Verlag (Francfort), 2018.

Dubout K. and Schlagdenhauffen R. Une archive inédite : le Journal intime d'Eugène Wilhelm (1866-1951). *Le Magasin du XIXe Siècle*. 2014 ;4:274-276.

Gibbs F., Owen T. Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. Digital humanties quaterly. 2012 ;6(2). http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html

Massot M.-L., Sforzini A. and Ventresque V. Transcrire les fiches de lecture de Michel Foucault avec le logiciel Transkribus: compte rendu des tests. 2018. 2018. https://hal.archives-ouvertes.fr/hal-01794139v2

Michel C. and Trognot G. L'expérience utilisateur au cœur de la stratégie. *I2D – Information, données & documents*. 2015;53(4) :40-41. https://doi.org/10.3917/i2d.154.0040

Mounier P. *Les humanités numériques.* FMSH eds. (Paris), 2018.

Muehlberger G., et. al. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*. 2018 ;75(5):954–976 https://doi.org/10.1108/JD-07-2018-0114

Perrin E. Bulliot, « Bibracte et moi. Transcription collaborative des archives archéologiques du site de Bibracte. 2019. https://f.hypotheses.org/wp-content/blogs.dir/7177/files/2019/11/Tutoriel_Transkribus_V2.pdf

Schantz H. The history of OCR, optical character recognition. *Manchester Center, Recognition Technologies Users Association*, Manchester, United Kingdom, 1982.

Schlagdenhauffen R. Une écriture du désir bisexuel est-elle possible ?. *Langage et Société*. 2014;148 :53-73.

Schlagdenhauffen R. Retour sur une controverse franco-allemande : l'Affaire Paris-Berlin (1904-1914). *In* González Bernaldo P. and Hilaire-Peréz L. (eds.), *Les savoirs-mondes. Mobilités et circulation des savoirs depuis le Moyen Âge.* Presses Universitaires de Rennes (Rennes), 2015:109-117.

Soudan C. Introduction. *Les Dossiers du Grihl, Faire une édition numérique savante et critique en TEI de manuscrits du XVIIe siècle*. 2012. http://journals.openedition.org/dossiersgrihl/5411

Warwick C. Studying users in digital humanities. In Warwick C. et al. *Digital Humanities in Practice*, Cambridge University Press (Cambridge), 2012:1-22. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.305.6694&rep=rep1&type=pdf