GrEma: an HTR model for automated transcriptions of the Girifalco asylum's medical records

Grazia Serratore[1]**, Emanuela Nicole Donato**[2]**,**
Erika Pasceri[3]**, Antonietta Folino**[4]**, Maria Teresa Chiaravalloti**[5]

[1]University of Calabria/
Institute of Informatics and Telematics of the National Research Council (Rende, Italy),
grazia.serratore@iit.cnr.it
[2]Institute of Informatics and Telematics of the National Research Council (Rende, Italy),
emanuela.donato@iit.cnr.it
[3]University of Calabria (Rende, Italy), erika.pasceri@unical.it
[4]University of Calabria (Rende, Italy), antonietta.folino@unical.it
[5] Institute of Informatics and Telematics of the National Research Council (Rende, Italy),
maria.chiaravalloti@iit.cnr.it

Abstract

This paper deals with the digitization and transcription of medical records from the historical archive of the former psychiatric hospital of Girifalco (Catanzaro, Italy). The digitization is carried out in the premises where the asylum once stood and the historical archive is stored. Using the ScanSnap SV600 flatbed scanner, a copy compliant with the original for each document contained within the medical records is returned. Subsequently the different training phases of a Handwritten Text Recognition model with the Transkribus tool are presented. The transcription aims to obtain texts in an interoperable format, and it was applied exclusively to the clinical documents, such as the informative form, the nosological table and the clinical diary. This paper describes the training phases of a customized model for medical record transcription, named GrEma, presenting its benefits, limitations and possible future applications. This work was carried out ensuring compliance with current legislation on the protection of personal data. It also highlights the importance of digitization and transcription for the recovery and preservation of historical archives from former psychiatric institutions, ensuring these valuable documents remain accessible for future research and potential users.

1 Introduction

The historical archives of former psychiatric hospitals represent a cultural written heritage of inestimable value for different research areas. They are multidisciplinary resources, providing a comprehensive insight into the history of psychiatry, the role of asylums in mental illness treatment, and the evolving cultural and social perception of mental illness over time.

Following the entering into force of the Basaglia Law in 1978, which mandated the definitive closure of asylums in Italy, there has been a growing recognition of the importance of studying these archives; among them the historical archive of the former psychiatric hospital of Girifalco (Catanzaro, Italy) attracted the interest of many scholars.

In Southern Italy, before the official opening of the Girifalco asylum, patients were systematically hospitalized in the Royal asylum of Aversa (Caserta) until, in July 1877, new admissions were barred due to a lack of space. Consequently, most patients in the province of Catanzaro were kept in private homes under the custody of family members and friends. Initially, the Provincial Deputation attempted to distribute the hospitalizations across the national territory, but due to considerable logistical and administrative difficulties, it was necessary to establish a new asylum in the Calabria region. As a result, in 1878, the Provincial Deputation approved the city of Girifalco as the site for new asylum and the decree for its opening was issued in 1880 (Greco, 2018).

The Girifalco asylum officially opened in 1881, and it became a point of reference for patients suffering from mental disorders, not only coming from the region. In almost a century of activity, 22,415 hospitalizations were recorded. However, today, the historical archive contains 15,794 medical records. The discrepancy between the number of hospitalizations and the number of medical records in the historical archive is due to multiple admissions of the same patient, as well as the loss and deterioration of some documents before their current arrangement (Chiaravalloti & Taverniti, 2021).

The medical records stored in the historic building of the former psychiatric hospital of Girifalco contain various documents, produced not only for clinical purposes but also for administrative and socio-demographic reasons, effectively representing patients' personal files.

All these documents are handwritten texts in Italian dating back to the 19th and 20th centuries.

Users seeking access to the content of a medical record in this archive must first obtain authorization from Archival Superintendence and subsequently visit the archive in person to consult the medical record. Nonetheless the documents could be not always fully understandable, as the legibility of the medical records may potentially be affected by their preservation status and the handwriting style.

Based on the access requests received so far, possible stakeholders and users interested in accessing the medical records are primarily scholars seeking to analyze their contents for various research purposes. There have been genealogical studies on the incidence of neurodegenerative diseases (Borrello et al., 2016; Cupidi et al., 2017; Bruno et al., 2022), examinations of how the lexicon and language used by clinicians and patients have evolved over time (Maria Teresa Chiaravalloti et al., 2020; Taverniti et al., 2023), reconstructions of the psychiatric history and the activity of this mental institution (Greco, 2018, 2020), as well as various anthropological studies (Costa & Serra, 2022). Additionally, some users may be interested in exploring their clinical family history or understanding the reasons why one of their ancestors or relative was institutionalized.

However, paper-based medical records can be difficult to consult because they require an on-site visit or because handwriting may be illegible. Moreover, frequent handling can accelerate their deterioration. Therefore, to make accessible the knowledge conveyed in these documents, digitization and transcription are two crucial processes that can improve the ability to analyze and interpret this cultural written heritage. In the archival domain, a good way to overcome barriers and issues related to document consultation and their accessibility is the adoption of innovative technologies for information retrieval in a digital environment. Digitization and transcription can be key processes to increase the accessibility and usability of documents (Jaillant, 2022).

The aim of this work is therefore to facilitate and promote the accessibility of this historical archive by providing authorized users with a digital and machine-readable format of the medical records.

This paper describes the digitization and transcription processes carried out on the medical records of the former psychiatric hospital in Girifalco, with a focus on the training phase of a Handwritten Text Recognition (HTR) model using Transkribus[1], a tool which allows automated text recognition and transcription.

The main aim is to present an automatic text recognition model that could improve the intelligibility and the interpretability of these historical documents. In detail, this paper is structured as follows: section II provides an overview of the types of data and documents considered. Section III outlines the techniques generally employed to preserve and protect archival heritage, focusing then on the digitization and transcription processes implemented to recover the historical archive of the Girifalco asylum. Finally, section IV details the training phases of GrEma, our HTR model for medical records, and presents the achieved results.

2 Data and documents

In Italy, the growing interest in the history of psychiatric institutions has led to several initiatives aimed at improving access to their historical archive. Many projects have been carried out to enhance the access to their written cultural heritage

[1] https://www.transkribus.org/

(Giuntini, 2009; Panattoni, 2009; Carrino & Di Costanzo, 2011; Milazzo 2020); among these, the General Directorate for Archives of the Italian Ministry of Culture promoted the project *Carte da legare*[2], which proposes an organic vision for the protection of the archival heritage of former psychiatric hospitals by surveying, reorganizing, and enhancing their archives (Kolega 2002).

The Girifalco asylum's archive is partially inventoried in *Carte da legare* thanks to the metadata reconciliation carried out by the Institute of Informatics and Telematics of the National Research Council (IIT-CNR), through the project ALPHA (eAsy InteLligent service Platform for Healthy Ageing). It involved the digitization of over 5,000 medical records, covering the period 1881-1931, and the manual transcriptions – through a dictation software – of the first 540 in chronological order.

For the purposes of this work, clinical documents were taken into consideration, particularly the nosological table, the informative form and the clinical diary.

The nosological table, compiled by the psychiatrist, includes personal, socio-demographic, and clinical information – such as etiology, diagnosis, hospitalization outcome, and medical observations. The informative form instead collects the patient's and family's health history at the time of the admission. Finally, the clinical diary consists of a set of notes relating to the patient's conditions during the hospital stay.

All medical records are handwritten in Italian and present syntactic and lexical characteristics, related to the psychiatric domain and the influence of the local dialect, that could make their interpretation more challenging.

3 Archive accessibility

The Italian Code for Cultural Heritage and Landscape, Article 6 (1), states that the enhancement of cultural heritage consists in the exercise of functions and activities aimed at promoting its knowledge and ensuring the best conditions for its use and public enjoyment (Code of Cultural Assets - Legislative Decree January 22, 2004, No. 42).

Enhancing accessibility is particularly crucial for historical archives of former psychiatric hospitals, as it is essential for preserving and valuing this collective memory while also fostering research on the documents. The digitization of the historical archive of the Girifalco asylum is therefore necessary to allow easier access without the need to directly consult the original documents. Digitization involves converting analog archival materials into digital format using specialized acquisition technologies. However, a proper digitization process of an archival document requires more than just a photographic acquisition. It is essential to provide an accurate transcription of every word to ensure the content is available in an interoperable format suitable for analysis.

3.1 Digitization

Digitizing historical archives provides several benefits but also presents challenges that must be carefully considered. Digitization creates faithful copies of the originals that can be easily stored and retrieved. Authorized users can consult digital copies remotely, facilitating access for researchers, scholars, and a broader audience. Monitored accessibility could offer logistical advantages, expanding the dissemination and appreciation of this archival heritage while also preserving the confidentiality of the data contained within the medical records.

The digitization of the historical archive of the former psychiatric hospital of Girifalco falls within the framework of the PRIN 2022 PNRR P2022R5LJ7 project, "Digital preservation, Linguistic analysis, and valorization of the historical archive of the former psychiatric hospital of Girifalco (DILIGO)", which involves the acquisition of about 3,000 medical records from the former psychiatric hospital dating from 1932 to 1944.

Following the authorizations granted by the Archival Superintendence, an agreement was reached between IIT-CNR and the ASP (Provincial Health Authority) of Catanzaro. It granted access to the historical archive of the Girifalco Asylum, that is located and stored in the historical building of this institution. The digitization was carried out on-

[2] https://cartedalegare.cultura.gov.it/home

site to ensure the security and preservation of the original documents.

Digitization encompassed the complete conversion of all medical records into digital format, as every document in them was scanned and digitized.

A preliminary organizational task focused on improving acquisition techniques by configuring the scanner settings to ensure proper image capture and the chosen device was the ScanSnap SV600 flatbed scanner, which offers a maximum resolution of 600 dpi. The documents were handled with gloves and were scanned on meticulously sanitized work surfaces.

Particular care was taken in handling fragile documents to prevent physical damage during scanning and to avoid any action that could further compromise their integrity. Unfortunately, due to their deterioration, some documents were already lost before our intervention, leaving inevitable gaps in the archive.

Decisions about acquisition settings were therefore taken by considering both the condition of the archival documents and the storage space required for the large volume of images. A medium-high quality setting was chosen, allowing file compression during acquisition to maintain good image quality while minimizing storage needs.

A particularly delicate phase of the digitization is post-processing, which is essential for producing legible images and optimizing the subsequent transcription process. Post-processing may include image enhancement to facilitate text recognition, such as noise reduction or distortion removal, as well as binarization, which converts a color or grayscale image into a two-tone (black and white) image.

Several file format options were evaluated for the scanned images, with particular attention to ensuring document security and long-term preservation. Finally, the PDF (Portable Document Format) was chosen because it preserves the original appearance and content of the document, thereby reducing the risk of accidental or intentional alterations during transfer or sharing. PDF files can be optimized to reduce file size while maintaining high quality, and they are ideal for archiving due to their data integrity, security features, and the ability to embed specific metadata (Annex 2 - File formats and data migration. Guidelines on the creation, management, and preservation of electronic documents, 2020).

However, digitization also posed some challenges. One major difficulty was the need to perform the digitization on-site, which can significantly slow down the overall workflow, especially when dealing with a large archive, such as the one from the former psychiatric hospital of Girifalco. The reasons for this slowdown stem from several factors: first, obtaining the necessary access permits may require a long waiting period. Then, there is the need to allocate time for daily travel to the archive for each workday. Additionally, it is essential to bring all the necessary equipment for on-site digitization, and ample space must be arranged to set up a suitable workspace for the process. In our case, it was not possible to leave the equipment at the archive premises, so it must be disassembled and reassembled each day, adding to the overall time and effort required.

Another relevant challenge was finding a balance between file compression and image quality. This added an extra layer of complexity, as we had to carefully consider the file formats and compression methods to maintain both the quality and longevity of the documents for future use.

Finally, another important aspect is the accessibility of the digitized medical records.

Without proper metadata management and organization, documents and their contents may be difficult to search and use, thereby reducing their value as research resources. In this context, establishing a minimum set of mandatory metadata, adhering to international standards, will be a necessary step to ensure the information remains interoperable.

The digitization of the historical archive of Girifalco asylum must be approached with an awareness of its limitations, including the time required for digitization, image quality, and information accessibility. The benefits of this process can only be maximized through careful planning and the adoption of appropriate technologies.

3.2 Transcription

Transcription plays a fundamental role in improving the accessibility and usability of historical documents, especially for handwritten materials, where the presence of different handwriting styles and the condition of the paper can significantly impair the comprehension of the content.

Depending on document types considered, their amount, and the purposes of the transcription, different methodologies and tools can be adopted. Generally, three main transcription methods can be identified: manual, manually assisted, and automated transcription. All of them aim to convert handwritten text into a digital and machine-readable version.

The manual transcription requires careful reading and interpretation of the original manuscript while faithfully preserving the content, structure, and, if necessary, specific graphic or stylistic elements. This method ensures a high level of accuracy, but it is onerous in terms of time and human resources required.

Manually assisted transcription, on the other hand, involves the use of computerized tools to facilitate the process, with human intervention to supervise the final product. This kind of transcription may involve the use of voice dictation software. An operator reads aloud the content of a document, while speech recognition software converts the spoken words into written text. These tools can reduce the time needed for transcription but require a final review to ensure that there are no comprehension errors in the dictation. In this case, transcription issues may arise due to pronunciation, the use of specialized or technical terms that are difficult to understand, or the failure to recognize proper punctuation in sentence flow.

To facilitate transcription and reduce the human effort required, the use of OCR (Optical Character Recognition) and HTR tools has become increasingly widespread. These techniques rely on the ability to recognize and correctly associate the characters within a word automatically, speeding up the process. For this reason, they are particularly suitable for working with a large volume of documents. Specifically, OCR technology detects and identifies the characters in a digital image of a printed text, converting them into digital characters (character encoding) so that they can be read and processed by a computer. It is particularly effective with printed texts that use standardized and well-defined fonts. For handwritten texts, instead, characters are not reproduced in a standardized manner but vary significantly depending on the handwriting style and the historical period of the documents.

For the purpose of this work, various tools were initially evaluated to find the best compromise between time efficiency and the accuracy and reliability of the medical records transcriptions.

At first, the efficiency of various voice transcription software was assessed, including Web Speech API[3], Microsoft Dictation[4], Dictation.io[5], and Dragon v5[6] (Matheson, 2007). In particular, the latter was used in the above-mentioned ALPHA project for the transcription of the first 540 medical records of the former psychiatric hospital of Girifalco. However, the analyzed software did not prove to be an effective support for interpreting content, understanding unclear passages within the texts, or optimizing transcription time, as they require just reading and dictating the medical records 'texts.

After an evaluation of transcription times and the total amount of medical records to be processed, HTR was chosen as the preferred approach. The main software for HTR are Transkribus and eScriptorium[7].

The first is a consumer-level automated text recognition platform. On the other hand, eScriptorium is an open-source tool that can be freely installed on a local machine and ca be used offline, but it offers fewer features and has a less user-friendly interface. After careful evaluation, due to its ease of use and the availability of comprehensive online documentation and support, Transkribus appeared to be the most suitable solution for achieving a balanced compromise between transcription accuracy and execution speed.

3.2.1 Transkribus

Transkribus is a software for automated image-to-text recognition, broadening access to historical

[3] https://www.google.com/intl/it/chrome/demos/speech.html

[4] https://apps.microsoft.com/detail/9pfv3vp2hz7b?hl=it-IT&gl=IT

[5] https://dictation.io/

[6] https://www.nuance.com/dragon.html?srsltid=AfmBOorKxIwMdkHLsZ-8IjYtmfsdaBxUyj6kMc4cge5ndPfQ2eUlD-XE

[7] https://github.com/UB-Mannheim/escriptorium

collections (Nockels et al., 2025). It is based on Java and leverages deep neural networks to recognize and transcribe text (Spina, 2023). This platform today includes a large community of users who access the web application through a system based on credits (Muehlberger et al., 2019). It allows the creation of custom recognition models for a specific dataset. This tool requires an initial phase of training, and the accuracy of the resulting transcription depends on the quality of the images, the number of different handwriting styles, and the size of the training dataset.

Initially, it is necessary to create the ground truth of the model, that is the dataset from which it will learn to recognize the text, and then to train the model itself. This step may take time, but it is required to obtain better results, as it will subsequently enable faster processing of the documents of interest, as well as the simultaneous handling of multiple documents. Human intervention is always advisable at the end of the transcription process to double-check for potential recognition errors. Nevertheless, the revision workload is significantly reduced compared to the time required for a manual transcription.

The platform offers pre-trained models and super-models shared by the community for various languages, but users can also request the training of customized models. In fact, depending on the specificities of the document types to analyze and their time and place of creation, it may be necessary to build a customized model.

The process of creating a customized HTR model is iterative and involves the progressive adaptation of the algorithm to the training data for obtaining accurate transcriptions. This learning process is supervised and relies on labeled datasets, known as ground truth and consisting of text line images paired with their transcription, in order to learn how to recognize different characters and correctly match characters in the image with those in the text. The training dataset must be adequately representative of the various types of documents considered, including diversity in layout and handwriting styles.

Although this process may seem complex, Transkribus interface simplifies users' interaction by limiting the setup to a few key parameters.

The time required to train a model ranges from a few hours to several days, depending on the training dataset size and the computational infrastructure load. The result of the training process is a model capable of recognizing handwritten or printed documents similar to those present in the ground truth.

A part of the dataset is randomly selected as validation set, allowing an assessment of the potential accuracy of the transcriptions that can be achieved. In the best cases, the platform can produce automatic transcriptions of handwritten materials with a Character Error Rate (CER) below 5%, meaning that 95% of the characters are correctly recognized, and between 1% and 2% for printed texts. However, results are considered optimal when the CER does not exceed 20%. If this threshold is surpassed, automatic transcriptions become less useful, as correcting numerous errors becomes more time-consuming than using other transcription methods.

Thanks to its flexible architecture, user-friendly interface, and ongoing development, Transkribus is increasingly establishing itself as a key reference point for the application of text recognition technologies, representing a valuable resource for archival and historical sciences.

4 GrEma: a transcription model for medical records

This paragraph aims to present the work phases carried out to train an HTR model for the medical records of the former psychiatric hospital of Girifalco, using Transkribus.

The first phase involved the construction of the ground truth. Considering the different types of documents present within the medical records, it was decided to transcribe only clinical documents, as they represent a valuable source of knowledge for countless future research. Specifically, the transcription was focused on the following clinical documents: i) the nosological table, ii) the informative form, iii) the clinical diary, and iv) the patient correspondence.

Considering the sensitive nature of the data processed, particular attention has been paid to the protection of personal data contained within these documents, in order to ensure compliance with current legislation. In fact, despite Transkribus privacy policies stating that documents are stored on READ COOP servers in compliance with the European General Data Protection Regulation (GDPR), it was decided to avoid uploading to the platform all administrative documents, cover pages of the medical records and documents containing personal data capable of making a patient or one of his/her family members identifiable.

Being aware of the efforts that would be required to train a customized model, it was at first decided to test one of the already available public models on Transkribus, in order to evaluate its performance on Girifalco's medical records.

In particular, an attempt was made with one of the most developed public models for the Italian language, which is Transkribus Italian Handwriting M1. This model is specifically designed for handwritten Italian text from the 16th to the 19th century and has a CER of 6.70%[8].

While model M1 is designed for documents from a period close to that of the medical records used in this work, it did not perform optimally on these specific documents. The factors that likely compromised its effectiveness were the peculiar handwriting styles, the specialized vocabulary related to the psychiatric domain, and the page layout.

Consequently, it was decided to train a specific model for the medical records. We decided to call it GrEma from the conjunction of the names of the researchers who developed it.

Generally, carrying out an effectively handwritten text recognition process requires around 15,000 transcribed words (approximately 75 pages), while printed text requires around 5,000 words (about 25 pages). For this reason, it was initially necessary to ensure that we had a ground truth composed of a representative number of images of the medical records and their corresponding transcriptions.

The first training session was conducted with a total of 20,776 words (approximately 130 pages), while the final dataset used to train the model consists of 94,624 words.

GrEma was trained using Pylaia as its engine, which is based on PyTorch. A 10% portion of the original dataset was selected as validation set, as this was considered the most suitable option given the relatively limited number of transcribed pages. GrEma used the public model M1 as its pre-existing base to leverage already available data for improved transcription performance. In fact, Transkribus offers a feature that allows users, during the training of their own custom model, to build upon existing public models. This functionality enhances the learning process by leveraging previously trained data, leading to improved performance.

After the first training session, the CER was 19.90%. It decreased to 16.92% in the second training phase with a dataset of approximately 34,500 words. Further training reduced the CER to 14.70 %.

GrEma was subsequently trained, achieving additional improvements in the CER value, which settled at 14.04%. In this case, the Word Error Rate (WER) was 39.03%. This metric indicates the percentage of words in the automatic transcription that do not match those in the ground truth transcription. However, WER tends to be higher than the CER, because a word is considered incorrect even if it differs from the reference by just a single character. As a result, WER may not always provide a fully reliable or representative measure of the model's performance, as such discrepancies do not necessarily make the transcription unintelligible, although manual correction is still required to ensure full accuracy.

All the training phases and the related results obtained are summarized in Table 1 and Fig. 1.

Training phase ID	**Training set size (pages)**	**Number of words**	**CER***
ID1	129	20,776	19.90%
ID2	245	34,512	16.92%
ID3	347	46,621	14.70%
ID4	566	94,624	14.04%

Table 1: Training results of GrEma for the transcription of medical records of the former psychiatric hospital of Girifalco.

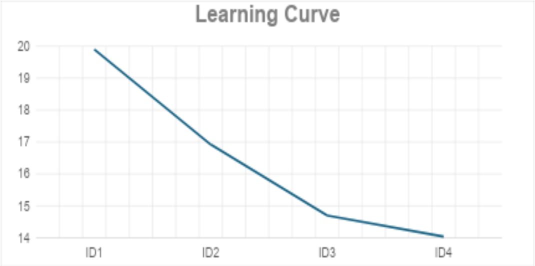


Fig. 1: Learning curve of GrEma during the training phases.

The learning curve in Fig. 1 shows the trend of the CER throughout the different training phases. The x-axis displays the ID corresponding to each of the four training phases of our model, while the y-axis represents the CER percentage achieved in

[8] https://app.transkribus.org/models/text/38440

each of them. The CER decreases progressively over the course of the training process, reflecting the diminishing returns in performance improvement as the dataset size increases.

When training a model with Transkribus, once the model achieves a relatively low CER, further reducing errors become increasingly challenging. This is because the model has already learned most of the patterns from the data, and additional improvements require a significantly larger amount of training data.

In Figure 2, it is shown how the CER is also influenced by the progressive increase in the number of training epochs: as the number of epochs increases, the CER decreases.

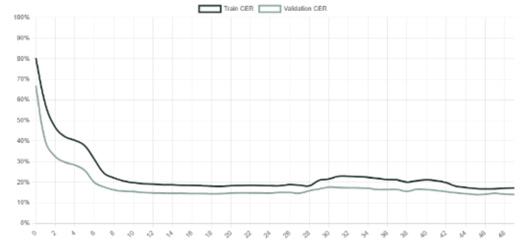
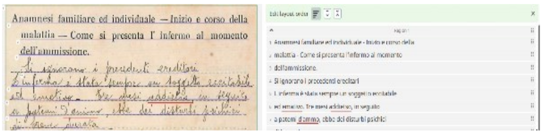


Fig. 2: The training chart shows how the CER value changes as the number of epochs for the GrEma model increases. The x-axis represents the number of epochs, while the y-axis indicates the CER percentage.

Another important aspect to take into consideration is that when a model is used to make a transcription, the output is not a direct transcription but rather a confidence matrix that assigns probabilities to the presence of each character in a specific position within the text. Consequently, the output is not influenced by the grammatical structure or syntax of the target language. The model may assign high probability to characters that visually resemble the original handwriting but may not form meaningful words or follow the rules of the language. This is because the model focuses on character recognition rather than understanding the context or meaning of the text.

The character-based probability approach also explains why GrEma often struggles with visually similar characters. For instance, the confusion between *u* and *n*, *o* and *a*, or *s* and *r* is common. These errors occur because the model relies solely on visual patterns rather than contextual understanding. Such mistakes are not always present but sometimes tend to appear in context where even a human reader might have difficulty in distinguishing the characters.

As shown in Fig. 3, GrEma produced a good transcription of the medical records, but still containing some kinds of errors.



Anamnesi familiare ed individuale — Inizio e corso della malattia — Come si presenta l' infermo al momento dell'ammissione.

Si ignorano i precedenti ereditari
L'inferma è stata sempre un soggetto eccitabile
ed emotivo. Tre mesi addietro, in seguito
a patemi d'animo, ebbe dei disturbi psichici
di breve durata

1 Anamnesi familiare ed individuale - Inizio e corso della
2 malattia - Come si presenta l'infermo al momento
3 dell'ammissione.
4 Si ignorano i precedenti ereditari
5 L'inferma è stata sempre un soggetto eccitabile
6 ed ematvo. Tre mesi addietso, in seguito
7 a patemi d'ammo, ebbe dei disturbi psichici

Fig. 3: An example of transcription made with GrEma, with CER at 14.04%.

In this example, the word “*emotivo*” is incorrectly transcribed as “*emativo*”; “*addietro*” appears as “*addietso*”; and “*d'animo”* is transcribed as “*d'ammo*”. In particular, the last two errors are unlikely to occur with human transcribers, whose syntactic knowledge of the language would allow them to infer the correct word even when not every letter is clearly legible. In addition, human transcribers would be aware of which sequences of letters are grammatically acceptable in the target language.

Therefore, additional post-processing is often necessary to validate the model's output and produce more accurate and readable transcriptions. In the future, this step could be crucial for fixing character misrecognitions and refining the text, especially when the model struggles with complex handwriting styles. However, identifying errors in certain transcription contexts opens up the possibility of automating the correction process by defining fixed rules, for instance with Python. Using regular expressions, dictionary-based checks, and language models might be useful to create automated scripts that detect common misrecognitions and apply predefined corrections, reducing the need for manual intervention and improving the overall efficiency and accuracy of the transcription process.

However, recently we trained our model again, but despite using a training set composed of 130,715 words, the CER did not decrease as expected but instead rose to 18.23%. An explanation for this increased value can be attributed to the inclusion of new handwriting styles in the analyzed dataset. The introduction of new handwriting styles remains a primary adaptation challenge for a model. Nonetheless, as medical records have been written by different hands over the years, it is not uncommon to find multiple handwriting styles even within the same document, as often happens in the clinical diary, because it was used to document the patient's stay in the hospital, and the physician on duty was responsible for making the necessary entries.

For handwritten texts, optimal performance is achieved when the texts are written by the same hand. However, within the same historical period, similarities can be observed in the way certain characters are shaped. Consequently, even when different handwriting styles are present, these similarities can help the model recognize a character or a sequence of characters with high probability. For instance, at the end of the 19th century, it was common practice to use a single uppercase *S* to indicate two *s* in the middle of a word, and we found it often recurring into the Girifalco's medical records too.

However, if the validation set includes a handwriting style that is underrepresented in the training dataset, the CER will inevitably be higher, even though the model may still provide highly effective transcriptions. This highlights how, in some cases, the CER value may not accurately represent the model's final performance.

In addition to the challenges posed by the continuous integration of new handwritings into the model, another significant limitation of this work arises from the need to manually transcribe specific pages of medical records, in order to protect sensitive and personal data. This manual transcription is necessary to safeguard sensitive information, ensuring that no private data is inadvertently shown. At the same time, this process introduces additional human efforts, hindering the ability to benefit from a fully automated transcription process.

Despite these limitations, the use of automatic HTR systems remains promising, as it helps reduce transcription times (as Transkribus allows to upload on the platform several documents to be transcribed simultaneously) and aids in recovering document content. In fact, one significant advantage of automatic transcription is its ability to recognize words that could be challenging for a human interpreter to decipher. This capability allows for the retrieval and transcription of text portions that would otherwise be lost with traditional methods, especially when dealing with handwriting styles difficult to read.

5 Conclusion

This paper presents the digitization and the transcription activities realized to preserve the medical records of the former psychiatric hospital of Girifalco. In particular, it describes the training phases of a customized model for transcribing clinical documents, named GrEma, outlining the steps taken to build it and the results achieved.

The digitization of the historical archive of the former psychiatric hospital of Girifalco required particular attention to ensure that the documents were digitized, preserving their integrity. Initially, careful planning of the digitization process was required, as it was necessary to physically reach the location where the archive is stored in order to go on with the digitization. Furthermore, the scanning equipment could not remain in the premises of the archive, meaning that it had to be disassembled and reassembled each day, with the equipment being transported back and forth. To address these needs, the equipment was carefully selected for its portability, leading to the decision to use a portable scanner. Considering the large volume of documents to be digitized, it was crucial to carefully evaluate how to manage the long-term preservation and the storage space, particularly in terms of selecting the most suitable file format. Consequently, the PDF format was chosen, as it can be optimized to reduce size while maintaining high quality. PDFs are ideal for archiving due to their data integrity, security features, and the ability to ensure long-term preservation and accessibility.

As the digitization, the transcription presented some challenges, including the data protection aspects and the adaptation of the HTR model to the documents' peculiarities. To ensure GDPR compliance, it was decided not to process administrative documents or others containing sensitive information.

As concerns the documents characteristics, GrEma was trained with progressively larger datasets to improve its accuracy. However, while the CER decreased over multiple training phases, ultimately reaching 14.04%, integrating new handwriting styles led to unexpected increases in errors. Obstacles such as confusion between visually similar characters persisted, highlighting the need for post-processing to refine the results.

Despite these challenges, automatic transcription offers significant advantages. It reduces the time required for the transcription process, enhances data accessibility, and enables the recovery of text that might be difficult for human readers to decipher. Although transcribed texts are in natural and unstructured language, different Natural Language Processing techniques could be applied to analyze them in the future, exploring their contents, increasing inferences, and creating new research opportunities, for example, in the fields of linguistics, medical history, and neurodegenerative diseases.

In fact, the digitization and transcription of these medical records not only allow to preserve important historical data but have the potential to transform them into dynamic resources for multidisciplinary research, improving both their accessibility and usability providing new opportunities for research.

A key future direction of this work involves the development of a digital platform inspired by the model of the Cambridge Digital Collection Platform (CDCP) adopted by the Cambridge University. The idea is to make available on this platform all digitized medical records, according to the IIIF (International Image Interoperability Framework) standard, and their corresponding transcriptions, encoded in XML-TEI. This would allow authorized users to engage with enriched versions of the medical records, accessing both a faithful copy of the original documents and its structured textual transcription. The adoption of XML-TEI encoding would further enable users to navigate the internal structure of the documents and their contents according to their specific research needs. In order to guarantee proper archival treatment and contextualization of each document, it will be necessary to integrate international archival standards such as EAD (Encoded Archival Description) and ISAD (G) (General International Standard Archival Description) into the platform's architecture.

Additionally, it is also essential to continue the transferring of the medical records metadata in accordance with the *Carte da legare* project guidelines, which promote standardized cataloging of medical records and ensure the creation of a coherent and aggregated dataset that will allow for broader statistical analysis and cross-institutional research based on harmonized data.

The difficulties faced during this work underscore the complexity of transcription for historical documents, where factors like handwriting variation, specialized vocabulary, and document degradation pose significant barriers to accurate recognition. Nevertheless, it is essential to continue these efforts in order to preserve and enhance the medical records from the Girifalco asylum, fully exploiting and discovering the knowledge contained within them.

Acknowledgments

This work was supported by PRIN 2022 PNRR P2022R5LJ7 "DIgital preservation, LInguistic analysis and valorization of historical archive of the former psychiatric hospital of GirifalcO (DILIGO)", funded by the European Union - Next Generation EU, Mission 4 Component 1 with CUP B53D23030290001.

This study is also part of a PhD project cofunded by the National Research Council of Italy (CNR) institutes: CNR-ICAR (Institute for high performance computing and networking, CNR-ITM (Institute on Membrane Technology), CNR-IIT (Institute of Informatics and Telematics).

The work has been conceived as a unit by all co-authors, and they are jointly responsible for paragraphs 1 and 5. However, Maria Teresa Chiaravalloti is responsible for paragraph 2, while Antonietta Folino authored paragraph 3. Emanuela Donato oversees paragraph 3.1, Grazia Serratore paragraph 3.2, and Erika Pasceri paragraph 3.2.1. Finally, Emanuela Nicole Donato and Grazia Serratore share responsibility for paragraph 4.

References

Agency for Digital Italy (AGID). *Guidelines on the creation, management, and preservation of electronic documents*. Annex 2: File formats and data migration. 2020.

Luara Borrello, Chiara Cupidi, Valentina Laganà, Maria Anfossi, Maria Elena Conidi, Nicoletta Smirne, Maria Taverniti, Raffaele Guarasci, and Amalia Cecilia Bruni. 2016. Angela R.: A familial Alzheimer's disease case in the days of Auguste D. *Journal of Neurology*, *263*(12), 2494–2498. https://doi.org/10.1007/s00415-016-8294-x

Francesco Bruno, Valentina Laganà, Raffaele Di Lorenzo, Amalia Cecilia Bruni, and Raffaele Maletta. 2022. Calabria as a Genetic Isolate: A Model for the Study of Neurodegenerative Diseases. *Biomedicines*, 10(9). https://doi.org/10.3390/biomedicines10092288

Candida Carrino and Raffaele Di Costanzo. 2011. *Le Case dei Matti. L'archivio dell'ospedale psichiatrico "S. Maria Maddalena" di Aversa 1813-1999.* Filema Edizioni, Napoli.

Maria Teresa Chiaravalloti and Maria Taverniti. 2021. Sanus egredieris. *Mélanges de l'École française de Rome - Italie et Méditerranée modernes et contemporaines*, 133–1.

Domenico Costa and Raffaele Serra. 2022. *Mangiare da matti: Una storia socio-alimentare a Girifalco (e non solo)*. Progetto 2000.

Chiara Cupidi, Valentina Laganà, Nicoletta Smirne, and Amalia Cecilia Bruni. 2017. The role of historical medical archives in the genealogical rebuilding of large families affected by neurodegenerative diseases. *Journal of Neurology & Neuromedicine*, 2(5).

Legislative Decree No. 42. 2004. Code of Cultural Assets.

Andrea Giuntini, editor. 2009. *Povere menti. La cura della malattia mentale nella provincia di Modena fra Ottocento e Novecento*. Tipografia TEM Modena, Modena.

Oscar Greco. 2018. *I demoni del Mezzogiorno: Follia, pregiudizio e marginalità nel manicomio di Girifalco (1881-1921)*. Rubbettino, Soveria Mannelli.

Oscar Greco. 2020. Migration trauma and psychiatry in the early twentieth century. *Journal of Modern Italian Studies*, 25(5):620–644.

Lise Jaillant. 2022. *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*. 1st ed., Vol. 2. Bielefeld University Press/transcript Verlag, Bielefeld.

Alexandra Kolega. 2002. Carte da legare: il progetto per il recupero e la valorizzazione degli archivi degli ex ospedali psichiatrici. *Archivio trentino*, 51(2).

Maria Teresa Chiaravalloti, Maria Taverniti, and Francesca Maria Dovetto. 2020. *Le cartelle dell'ex ospedale psichiatrico di Girifalco. Lessico, strumenti e terapie*. Lingua e Patologia. I sistemi instabili. Aracne.

Jennifer L. Matheson. 2007. *The Voice Transcription Technique: Use of Voice Recognition Software to Transcribe Digital Interview Data in Qualitative Research. The Qualitative Report*, 12(4), 547–560.

Guenter Muehlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, Philip Kahle, Mario Kallio, Frederic Kaplan, Florian Kleber, Roger Labahn, Eva Maria Lang, Sören Laube, Gundram Leifert, Georgios Louloudis, Rory McNicholl, Jean-Luc Meunier, Johannes Michael, Elena Mühlbauer, Nathanael Philipp, Ioannis Pratikakis, Joan Puigcerver Pérez, Hannelore Putz, George Retsinas, Verónica Romero, Robert Sablatnig, Joan Andreu Sánchez, Philip Schofield, Giorgos Sfikas, Christian Sieber, Nikolaos Stamatopoulos, Tobias Strauß, Tamara Terbul, Alejandro Héctor Toselli, Berthold Ulreich, Mauricio Villegas, Enrique Vidal, Johanna Walcher, Max Weidemann, Herbert Wurster and Konstantinos Zagoris. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

Fabio Milazzo. 2020. *Una guerra di nervi. Soldati e medici nel manicomio di Racconigi (1909-1919)*. Pacini, Pisa.

Joseph Nockels, Paul Gooding, and Melissa Terras. 2025. Are Digital Humanities platforms facilitating sufficient diversity in research? A study of the Transkribus Scholarship Programme. *Digital Scholarship in the Humanities*, 40.

Riccardo Panattoni, editor. 2009. *Lo sguardo psichiatrico. Studi e materiali dalle cartelle cliniche tra Otto e Novecento*. Bruno Mondadori, Milano.

Salvatore Spina. 2023. Handwritten Text Recognition as a digital perspective of Archival Science. *AIDAinformazioni*, 1–2.

Maria Taverniti, Maria Teresa Chiaravalloti, and Francesca Maria Dovetto. 2023. *Uno sguardo sociolinguistico sui pazienti dell'OP di Girifalco*. Università di Napoli Federico II, Napoli.