Short-Cutting Manual Acquisition in Deep-Learning Deciphering of Old Documents

Dan Cristea[1,2], *Petru Rebeja*[2]

[1]Romanian Academy, Institute for Computer Science
[2]"Alexandru Ioan Cuza" of Iași, Faculty of Computer Science
dan.cristea@acadiasi.ro, petru.rebeja@info.uaic.ro

Abstract

We present an approach to ease the effort of acquiring annotation data intended to train a technology for automatic transcription and transliteration of old documents. The research is motivated by the interest to decipher in the Latin script Romanian documents written in Cyrillic along the centuries XVIth-XIXth. The whole enterprise is briefly described, then the attention concentrates on an alignment algorithm that reuses manual transcripts for the benefit of automatically acquiring training data to enhance neural network models. The approach is easily reproducible for other languages.

Index Terms: old documents deciphering, resource acquisition, image to text conversion, cultural heritage

1. Introduction

During centuries, the writing has changed a lot. This is even more evident in languages that have commuted from one script to another, as is Romanian. On the territory of the actual Romania a variant of the Cyrillic script was mostly used until the first decades of the XIXth century. Then, for a short period, a mixture between Cyrillic and Latin signs, called the transition alphabet, replaced the Cyrillic, until the Latin script was fully adopted. The need to recuperate, for research, study and the genuine leisure of reading of the large public, more than 2000 old Romanian documents, as inventoried in Romanian and foreign libraries, by Cândea [1] and Bianu et al. [2], to which manuscripts deposited in monasteries should be added, strongly motivates this research.

The documents we concentrate on include printings and uncials (documents handwritten by copyists reproducing shapes of original printed characters), originating from the centuries XVIth to XIXth. Transposed in digital form by scanning, a large part of these documents can then make the input to an interpretation and transliteration technology from Cyrillic to Latin fonts. The research reported in this paper is a follow-up of the DeLORo project (Deep Learning for Old Romanian – see Acknowledgments), whose main achievements have been: to develop technological solutions for locating lines and characters that appear in the scan of a page; to interpret in context Cyrillic characters; to assign Latin letters to them; to align character frames from page images with the corresponding decoded text in the case of documents that are interpretatively transcribed; to recompose a linear sequence of characters, including there where interlinear writings occur; to recompose words, also in cases when white spaces are partially missing; to guess lemmas and part of speeches of a subset of the inflected word forms belonging to the old Romanian language; and to judge the semantic contexts in which words are used – a decisive step towards the disambiguation of meanings.

This paper presents the general DeLORo enterprise, with a focus on the solution we adopted to speed up the process of acquisition of training data in a low-resourced language as Old Romanian. This section continues with a brief description of the data acquisition process in DeLORo and a very resuming presentation of some approaches that have similarities with ours. Section 2 then describes the model, Section 3 – the results and Section 4 presents some conclusions.

1.1. Acquisition of data in DeLORo

Most data used to train the deep learning models for character identification and recognition has been acquired in DeLORo by manual annotation, using a specially built front-end – OOCIAT (the Online Old Cyrillic Image Annotation Tool [3]). On the images of pages of old documents, fetched from ROCC (Romanian Old Cyrillic Corpus), the annotators were concentrated on two tasks: to frame in rectangles graphical objects (representing titles, lines of text, characters, modifiers, marginal strings, interlinear strings, initial letters, reference marks, frontispieces, ornaments, etc.) and to fill in content values for some of these objects (titles, lines, characters, etc.). Although the amount of annotation work has been considerable in the project, involving not only linguist experts but also less experimented personnel, in many cases the resources thus acquired proved not being enough to obtain high-quality deciphered objects through deep learning.

Figure 1 shows schematically how is stored a character in the ROCC database. Similar XML-like structures are designed to describe all types of graphical objects.

To speed up this acquisition process, we decided to use for training purposes also a collection of documents that display interpretative transcriptions of the original Cyrillic books. These transcriptions are either produced by experimented linguists, as critical editions, or are thesis of PhD students in Paleolinguistics.

```
<object:Character>
    @objectId <!-- (required)-->
    @objectAnnotator <!-- (required) annotator's user-name, or "AUTOMATIC" if the
        object is added by the machine -->
    @objectAccuracy <!-- (optional) if @objectAnnotator = "AUTOMATIC", this
        attribute records the trust level in its automatic identification -->
    @objectContent <!-- (opțional) when occuring: "VOID" => the character is not
        transcribed; 1 or 2 Latin letters otherwise -->
    <objectCoordinates> <!-- (required) the rectangle's coordinates encapsulating
        the character in the image -->
        @leftUpHoriz <!-- the horizontal coordinate of the upper left corner -->
        @leftUpVert <!-- the vertical coordinate of the upper left corner -->
        @rightDownHoriz <!-- the horizontal coordinate of the lower right corner-->
        @rightDownVert <!-- the vertical coordinate of the lower right corner-->
    </objectCoordinates>
</object:Character>
```

Figure 1: *Representation of an object of type character in ROCC*

1.2. Related work

Curation of old writings is a stringent preoccupation for revitalizing the cultural heritage of the past. Image interpretation and document processing have both reached performances that make them current technologies in many types of applications. Optical Character Recognition (OCR) technologies have become extremely performant for a high diversity of printed fonts and even for cursive handwriting. FineReader, for instance, which integrates intelligent document processing features, has become one of the major commercial products in content-centric processing. It combines machine learning, natural language processing and computer vision techniques to decipher and correctly interpret the content of documents. Another well-known application, Transkribus [1], also an ABBYY AI-based system, decodes handwriting and even *scripta continua*. However, the documentation does not suggest the capacity to recuperate and place in sequence interlinear writing, while tests with a Cyrillic writing produced incomprehensible Latin strings. The µDoc.tS platform [4] decodes handwritten Old Greek documents, being able to do keyword spotting in handwritten text. The platform helped making public a consistent set of handwritten documents originating from the Stavronikita Monastery on Mount Athos.

A classical OCR approach would prove ineffective for the tasks we envisage, especially because of the large diversity of types of objects to be identified in the scanned pages, and for the non-linear placement of these objects, each of them having a particular contribution to the overall understanding of the content.

2. The model

2.1. The general workflow

The general architecture of our approach is sketched in Figure 2. A laborious process of data acquisition prepares the data (stored in the ROCC database) for training the neural machinery that detects objects in the scanned images of pages and deciphers the content of some of these objects. Then, once trained, the model can be applied on new pages.

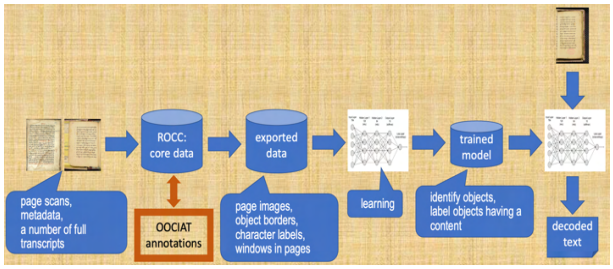
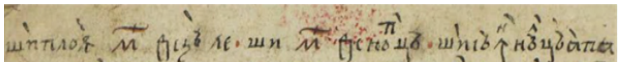


Figure 2: *Data flow in DeLORo*

The placement of the graphical objects within a page are indicated by coordinates of their left-up and right-down corners (see Figure 1) and no relative positions to each other is marked. This allows for a random order of annotation of the objects, freeing the annotators of any cumbersome order constrains in the annotation process. The sequencing of the text can be solved even if there is no explicit definition of the sequence of lines, by geometrical considerations of the position of the rectangles that frame them. The deduction of relative positions can be performed for many types of objects and is important in reconstructing the sequentiality of the text.

Figure 3 shows an example of a row of text which displays also interlinear writing and its transliteration, with notations that put in evidence the shift of attention in reading and the placement of white spaces.



ș i ▫ p l o ă 40 d e ▫ z î ☒ l e ș i 40 d e ▫ n o ↑ p ↓ ț i ș i ▫ s ă î ↑ n ↓ n ă ↑ l ↓ ț ă ▫ a p a

Figure 3: *An original line of text (from Hronograf, page 2, verso, DeLORo file 002v.jpg, approx. "and it rained 40 days and 40 nights and the water rose"), with the original line (up) and the transliterated one (down); the notations used here are: ↑ = the next character is placed above the line; ↓ = the reading continues in the line; ▫ = a space is missing; ☒ = a space should be removed.*

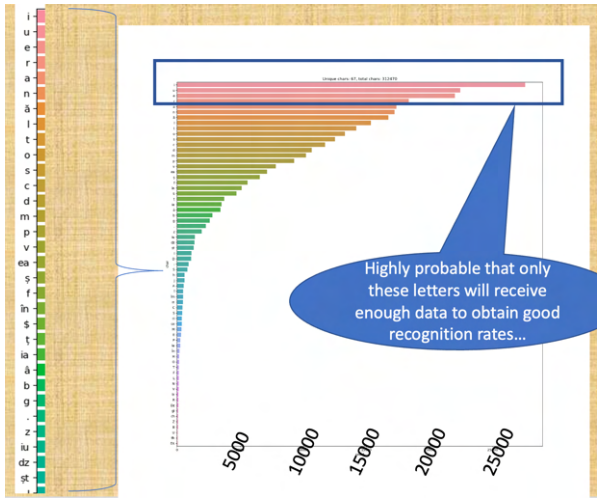


Figure 4: *Density of letters in old documents*

The more data we have, the better the training of the labeling module. However, Figure 4 shows that there is a high non-homogeneity in the density of letters of a Cyrillic-Romanian text. This situation cancels down the hope that a constant acquisition process would finally fulfill the data needs.

2.2. Preparatory operations

To overcome the scarcity of this type of data, we decided to use the small number of old books from the ROCC collection that contain not only images of pages but also the equivalent edited texts, therefore which are completely transliterated, either in critical editions, by linguist experts, or as PhD thesis in linguistics, by students. In this paper we show how could this precious data be used for training the deciphering neural technology. It is evident that the difficulty stays in the fact that an alignment between images of pages and the transliterated

[1] https://readcoop.eu/transkribus/?sc=Transkribus

text should be created beforehand tempting to use the image-to-text data in the training process.

The need to put objects of different shapes and types in order has led us to adopt a deciphering process that goes along three main steps: 1). objects identification (OI), 2). objects labeling (OL), and 3). objects sequencing (OS), in order to recover the sequentiality of the text. Both OI and OL are neural classification processes. The first one groups objects in large classes and the second attributes a content to those objects that may have a content, such as characters, for instance. The third step, OS, uses geometrical parameters and heuristics to place objects in sequences and to recover the string of words, by using the content of recognized objects (in case of a character, the content is its Latin label) and lexical information. As reported in [5] and [6], we obtained rather accurate character identification results, while the tests of labeling characters have proved much poorer.

As a prerequisite to this process, we have first created a section in the ROCC database that had to accommodate the <character image, interpreted text> alignments (see Figure 5).

```
<onePageAlignments> <!--(required): section automatically filled in by the OOCIAT editing
interface and the alignment program-->
    @pageID <!--(required)-->
    <seqAlignment> <!-- a record of this type is generated for each alignment between a
sequence of character frames and an string of characters in text, performed manually or
automatically-->
        @seqCharLength <!-- (required): length of the sequence -->
        @seqCharIds <!--(required): a sequence of <object:Character> IDs -->
        @seqTxt <!-- (required): the aligned text -->
        @goldTestAlignment <!-- (required): with the values: "gold" = alignment done
manually or by machine, then manually validated; "test" = automatically generated
alignment, not validated -->
    </seqAlignment>
</onePageAlignments>
```

Figure 5: *The sequence alignments section in the ROCC collection*

Then, all transliterated texts have been manually segmented by placing `<p>` markers to split the text in pages at the exact places in the .txt documents that separate two consecutive pages from the original scanned documents. As seen in Figure 5, the image part of each alignment in that section contains sequences of IDs of Cyrillic character objects in the image of the original page (`@seqCharIds`), which are actually frames of characters as identified by the object identification module, while the text part (`@seqTxt`) contains an equal length sequence of Latin characters. All alignments are considered at the level of a page.

The character frames in a page are linearized on rows and the position of rows is determined by computing the histograms of character boxes on a vertical axis of the page that goes downwards (see Figure 6). Then, positions of rows correspond to peaks of the histogram. This allows to compute approximate matches between all character boxes in a page and their textual transcription. As Figure 6 also suggests, only part of the boxes have a transcribed content, the others being void (with the meaning: not yet interpreted) and the idea is to fill in the missing parts based on the linear displacement of the full-content and empty boxes when comparing them with the transcribed text.

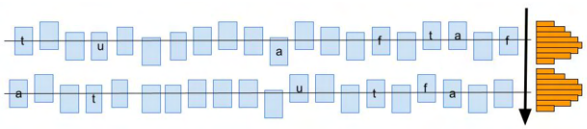


Figure 6: *Detecting lines*

2.3. Aligning character shapes to transcribed letters

The alignment algorithm should build only shapes-to-letters pairs that have very high (almost certain) recognition probabilities. Thus, the following heuristics is used to iteratively enlarge the set of characters recognized, with safe accuracy. Suppose that at the beginning of the iteration, a sufficiently high recall and precision of the OI module has been obtained and, based on the characters manually labelled, the OL module is already trained to recognize a small set of characters with a sufficiently high precision. Then:

a) the original page is segmented into character frames (see Figure 7, for just a squared window extracted from the page);

b) apply the initial model of the OL module to label all high confidence instances of this small set of characters in the page; now the image contains a small number of labelled character frames on lines and a much larger set of unidentified blank frames (like in Figure 8);

c) use the labelled character frames in the image as landmarks and count the blank frames in-between them; then, search the text for matching equivalent pigeonholes patterns on the lines of text;

d) select only short equal-length aligned sequences in the image and the text and include them in the ROCC alignment section as pairs of sequences of IDs of character frames and the strings of their labels;

e) do this over the whole collection of critical editions; using the alignments thus acquired, now the OL model can be trained on a larger set of instances for what beforehand were considered “unsafe” characters; retain only new labels which now pass the accuracy threshold considered “safe”;

f) iterate steps c) to e) until no new characters pass the accuracy test; indeed, the process must die out naturally, since more and more characters are now consumed as landmarks in each page.

We do not mark line borders on the transcribed text (only page borders, as mentioned before). As such, the alignments between character boxes and labels in the text should be deduced strictly based on the following clues: 1). the box offset in the string of boxes is an approximate match of the character offset in the text; 2). if the character box is decoded, then its label should be equal with the aligned character; 3). if the string of character boxes and the string of text characters on a page would be placed one above the other and alignment lines would be rendered between the boxes and their corresponding labels, the lines should not intersect. These restrictions suggest an alignment algorithm similar to those used to align translations [7], with the supplementary constraint that no intersections of alignment lines are allowed, therefore the image to text alignment matrix should be a strait diagonal.

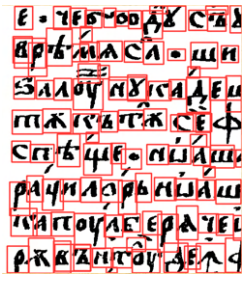


Figure 7: *Characters identified in a window by the OI module*

A sketch of this process is displayed in Figure 8. On the first two rows an initial segmentation of the Cyrillic characters in a line is shown and their expert Latin transliteration. The following two lines separate the characters whose confidence scores are above a certain threshold, qualifying them as "safe" and those under this threshold (there is no intersection of labels between these two classes). Finally, the last row puts in evidence the character boxes of the Cyrillic characters acquired through alignment.

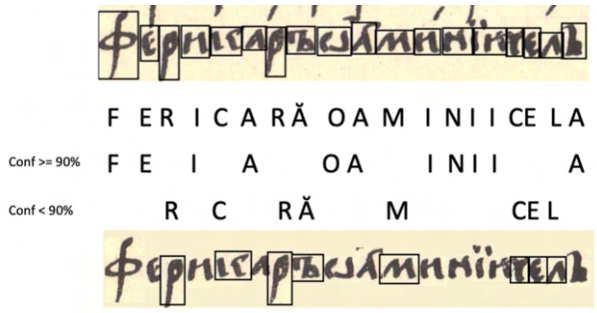


Figure 8: *Example on how the alignment strategy works*

The "short" distance mentioned at step d) in the heuristics sketched above (the limit was set to 5 in our implementation) is imposed in order to minimize the risk that a 1-to-2 ( , transcribed *gh*, as in  = *ghenadie*) or a 1-to-0 ( , transcribed void) equivalent between the Cyrillic and Latin transcription would invalidate the 1-to-1 alignment supposition. The chance to fall on an equal number of letters by first missing one and then adding one (or vice-versa) is small on small distances, but could be significant in large chunks. It should also be noted that the <image, text> alignment pairs in the ROCC collection are permanently updated during the iteration.

3. Results

The results and discussion in this section refer strictly to the short-cutting approach described here, which augments the data necessary to train the OL module after the manual acquisition process was stopped for lack of human resources, and not to how the OI and OL modules perform themselves. These modules have been evaluated in previous work (see [5] and [6]).

It is natural to consider that the behaviour noticed after processing just one page (recorded in the parallel format image + text) is statistically enough to infer almost similar results on the rest of the corpus. Therefore, we isolated a page displaying 775 characters. Out of these, only 702 characters have been automatically detected as character frames in the OI phase. The subset of characters for which the OL model originally reported sufficiently high confidence in evaluation (above 95%) comprised the letters: *u* (48 instances found on the test page), *t* (40 instances), *a* (27 instances) and *f* (12 instances), with a total of 127 occurrences. 9 strings of letters bounded by characters from the above set have been detected, all having in-between lengths of 1 to 5 characters. All were automatically identified, but 2 had to be ignored due to differences in the corresponding lengths (the string of character boxes in the image was not equal to the string of letters in the text), for example, because a final-word soft *i* was not transcribed or the OI model missed a character box.

Thus, the net profit of new alignments on the test page has been of 8 characters (about 1% of the total number of characters from the page). As our collection of imported data includes 25 fully transcribed documents, which sums up to approximately 6,300 pages, and considering an average of 700 characters per page, this leads to approx. 4,410,000 transcribed characters. Then, 1% means 44,100 new instances of characters acquired with this shortcutting strategy, only on the first iteration. Our total number of manually annotated character instances in ROCC was a little bit above 146,000 at the end of the two-years project. This means that the accumulated new data was approximately 3.3% of the originally acquired data, only on the first iteration. The 44,100 instances, distributed over the 27 letters of the Latin alphabet, were enough to raise the number of instances above the quantity level that would secure a trustful recognition for another two more letters, meaning that at the end of first iteration we gained two more character labels.

This are statistical data inferred on the whole corpus after counting the new labels whose training data have reached the "safe" threshold after just one iteration of the shortcutting algorithm. In reality, our runs have shown that repeating this process until no new acquisitions have been noted, has brought us for free data to safely train the OL module for 8 new character labels. This is supposed to have spared hundreds of hours of tedious annotation.

4. Conclusions

We presented an approach aimed to shortcut the tedious work of manual annotation intended to acquire the ground truth for neural networks automatic deciphering and interpretation in the Latin script of old Romanian documents originally drafted in Cyrillic. A large collection of images of old documents has been interactively annotated by members of the DeLORo project, but the amount of data proved insufficient for a trustful training of a deep learning model of character labeling. Our methodology puts to work also a sub-collection of documents for which the images of original pages are doubled by parallel transliteration of the textual content in the Latin script.

We show that exploiting the parallel image-text data and repeating an alignment-training-evaluation iteration the process of data acquisition can be bootstrapped up to the point to fulfil the model training task with much smaller quantities of manually acquired data. It is clear that the process has a natural exhaustion, because, as can be remarked from Figure 4, the least frequent letters will not have a sufficient number of instances in the entire collection of parallel data that can be used. For these low-frequency letters, the recognition should be based on linguistic grounds, in a similar manner in which people recognise characters, by using the context and their knowledge about the language.

The approach can be easily adapted for any under-resourced language.

5. Acknowledgments

The current work is a follow-up of the DeLORo project, PN-III-P2-2.1-PED-2019-3952, no. 400PED: "Deep Learning for Old Romanian", which has run between Oct. 2020-Oct. 2022.

We thank the members in the project from the Institute of Computer Science of the Romanian Academy, Iași branch, and from the Faculty of Mathematics and Computer Science of the University of Bucharest. Their names can be found in the project pages at http://deloro.iit.academiaromana-is.ro/.

We thank the Romanian Academy Library – for offering for research purposes their collection of scans of old books and the corresponding library metadata.

We thank the students in the Master of Paleolinguistics, Faculty of Letters from "Alexandru Ioan Cuza" University of Iași, which, guided by dr. Roxana Vieru, contributed with hundreds of hours of OOCIAT annotation.

6. References

[1] V. Cândea. "Mărturii românești peste hotare". Serie nouă. Vol. I-VI.1: Vol. I-IV, București, Editura Biblioteca Bucureștilor, 2011, 2012. Vol. V-VI.1, București, Editura Academiei Române, Editura Muzeului Literaturii, 2014, 2016.I.

[2] Bianu, N. Hodoș, and D. Simionescu, "Bibliografia românească veche. 1508–1830.". Tom. I–V, Ediţiunea Academiei Române, București, 2490 p., 1903-1944.

[3] D. Cristea, C. Pădurariu, P. Rebeja, A. Scutelnicu, M. Onofrei "Data Structure and acquisition in DeLoRo - A Technology for Deciphering Old-Cyrillic-Romanian document". Proceedings of the the 16th International Conference "Linguistic Resources and Tools for Processing The Romanian Language", ONLINE, 13-14 December 2021, ISSN 1843-911X, pp. 59-74, 2021.

[4] Tsochatzidis L, Symeonidis S, Papazoglou A, and Pratikakis I. (2021). HTR for Greek Historical Handwritten Documents. Journal of Imaging 7(12):260, 2021. https://doi.org/10.3390/jimaging7120260.

[5] D. Cristea, P. Rebeja, and C. Pădurariu. "Applying YOLOv5 Learning in Detecting Old Cyrillic Romanian Characters". In Svetlana Cojocaru, Victoria Bobicev, Tatiana Verlan, Dan Tufiș and Dan Cristea (eds.) Proceedings of the 17th International Conference "Linguistic Resources and Tools for Natural Language Processing", online, Chișinău, 10-12 November, "Alexandru Ioan Cuza" University Publishing House, ISSN 1843-911X, 2022, pp 115-122, 2022.

[6] D. Cristea, N. Cleju, P. Rebeja, G. Haja, E. Coman, A. Vasilescu, C. Marinescu, A. Dascălu. "Bringing the Old Writings Closer to Us: Deep Learning and Symbolic Methods in Deciphering Old Cyrillic Romanian Documents", in Memoirs of the Scientific Sections of the Romanian Academy, 2023 – to appear.

[7] M. Kay, M. Roscheisen. "Text-Translation Alignment", Computational Linguistics, vol. 19, no. 1, 1993.