OCR / HTR technologies and Armenian Heritage Preservation

Chahan Vidal-Gorène

DOI: https://doi.org/10.52027/18294685-cvo2023.sp

CHAHAN VIDAL-GORENE
Paris, France

OCR / HTR TECHNOLOGIES AND ARMENIAN HERITAGE PRESERVATION*

Keywords: digitization, handwritten text recognition, optical character recognition, Armenian archives, manuscripts.

Introduction

Character recognition is the task that automatically converts a scanned document into a searchable text format. A distinction is made between OCR (Optical Character Recognition) for printed documents and HTR (Handwritten Text Recognition) for handwritten documents. The objective of these steps is to provide a searchable and editable version of a document. These technologies meet the needs of institutions involved in the massive digitization of their collections, and strengthen the accessibility of documents and their preservation.

Although the issue of OCR for printed documents remains a task considered to be largely resolved, including for Armenian, with a wide variety of software such as Abbyy Fine Reader[2] (paid software), Tesseract[3] (free and open source) or Calfa[4] (paid service) that can reach a character error rate (CER) of less than 1%, the recognition of manuscripts or historical documents remains an open research problem. The great variability of handwriting, the philological questions inherent to the transcription of documents, the degradation of documents (damaged printed materials, old printed matter or damaged manuscripts), the inconsistency of layouts or the quality of digitizations are all issues to be overcome, and that limit the development of generic models of recognition.

Artificial Intelligence and pipeline for Armenian

With the rise of Artificial Intelligence however, these tasks can be overcome, provided an AI is trained with enough representative samples of the desired task. These samples, within the framework of an HTR, consist in documents manually transcribed and annotated according to specifications to be defined [Vidal-Gorène, 2023]. Many datasets have been compiled in recent years, particularly for Latin [Nikolaidou, 2022] and Arabic [Vidal-Gorène, 2021b] scripts, in order to overcome specific issues raised by scripts (e.g. Latin scripts from the 14th century) or layouts. Concomitantly several platforms have been developed to enable researchers to overcome the barrier of HTR on their corpus, the best-known being Transkribus [Kahle, 2017]. Nevertheless, these tools therefore require a large amount of data, the critical mass of which for Armenian is hard to reach, owing to linguistic specificities (e.g. variants of Armenian) or paleographic particularities (e.g. abbreviation system), and the variety of applications (e.g. ancient manuscripts, contemporary manuscript archives, etc.). Armenian is still today considered as an under-resourced language.

Calfa has implemented AI approaches to overcome the lack of data, by creating effective specialized layout analysis and recognition models with very few data [Vidal-Gorène, 2021a]. This results in recognition models that can achieve more than 99% good recognition for printed matter and more than 95% for manuscripts. For example, we have shown that on old manuscripts in bolorgir, our models prove to be more than 97% efficient with only 3 transcribed pages [Kindt, 2022], and for more cursive Arabic scripts more than 93% good recognition with 10 pages transcribed [Lucas, 2022][5]. These approaches are implemented in particular in our services and our annotation tool Calfa Vision (freemium), dedicated to non-Latin scripts and

* This paper is an excerpt from a couple of conferences on Digitization and OCR/HTR for Armenian documents that took place in fall 2022 in Armenia ("Heritage Preservation for a Sustainable Future" International Conference - National Library of Armenia, and Գիտաժողով՝ նվիրված Մայր Աթոռ Ս. Էջմիածնի «Վաչէ եւ Թամար Մանուկյան» - Etchmiadzin).

[2] https://pdf.abbyy.com/

[3] https://tesseract-ocr.github.io/

[4] https://calfa.fr

[5] Traditionally, similar approaches for printed documents require on average at least 100 transcribed pages [Ströbel, 2020].

poorly endowed languages [Vidal-Gorène, 2021a]. The pipeline implemented by Calfa is a joint work between machine and human, the expertise of the latter remaining essential [Vidal-Gorène, 2023].

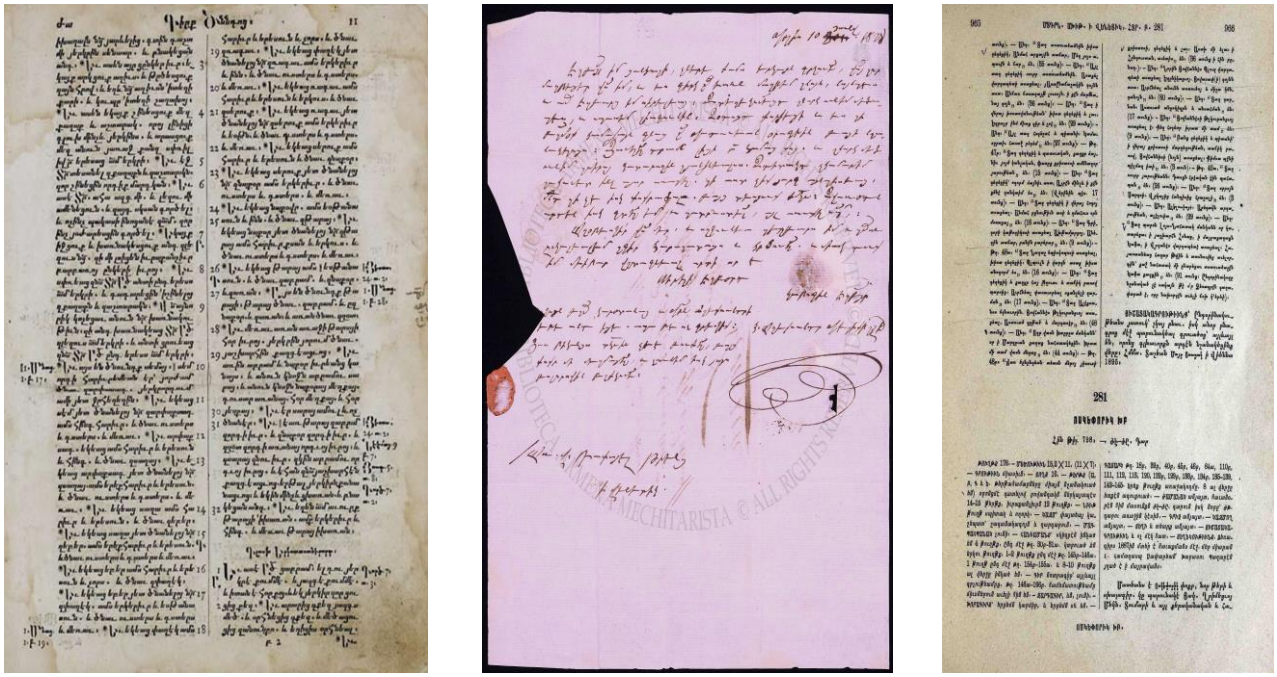
Some case studies and Results

Therefore, we achieve very specialized models (hence not as versatile) obtaining very high recognition rate on a given task (see Figure 1). We apply them to various case studies:

• old, poor quality or damaged digitization (handwritten or printed) requiring a dedicated recognition model;

• retrieval of specific information in a structured printed document;

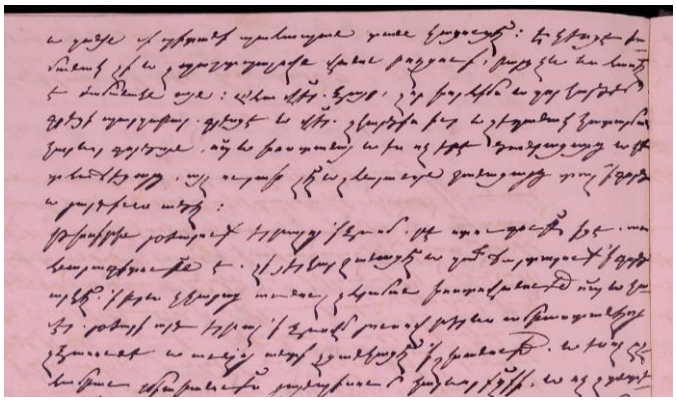
• ancient and modern manuscript documents;



• retrieval of information in an unstructured handwritten document.

Figure 1: Examples of documents targeted by specialized models. Left: Voskanian Bible (1666) with 98.67% of good recognition; center: handwritten letter from Father Trianz (1828) 98.9% good recognition; right: the catalog of Armenian manuscripts from San Lazzaro (Venice), whose average recognition is 99.5% with automatic content detection (manuscript title, number of pages, place of copy, etc.).

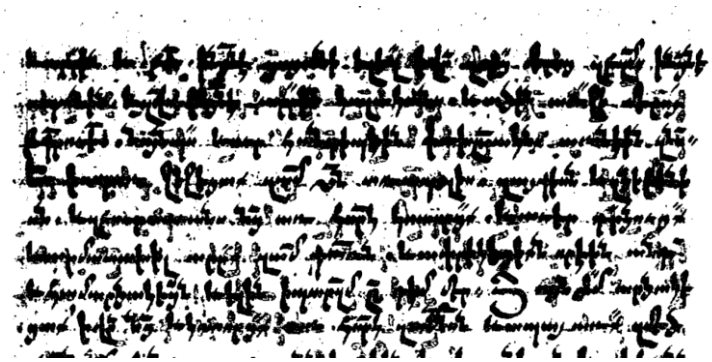
Calfa is notably involved in the processing of microfilms for the National Library of France [BnF Datalab, 2023]. The creation of a very specialized model on a microfilm of the BnF (P191) results in a 95.1% relevance. Although some parts of the microfilm remain illegible, the model manages to offer usable reading, with nonetheless frequent confusion between ա / ւ. Abbreviations have not been expanded by the model, and word separation is provided despite the fact the document looks like *scriptio continua* (see Figure 2).



և զանձն ի պիտանի պահապան տանն կացուցել։ Է կեսցէ ժա
մանակ զի և զպաշտպալովն վանսն թողցուք, բայց չև ևս հասել
է ժամանակն այն։ ահա վեր հայր, զոր խորհիմս և զոր կարծեմ
գրեցի պարզաբար. գրեսցէ և վր. վեր զկարծիս իւր և զեղանակ
կատարման
կարեւոր գոյերոյն, ոպ և խոստանայ և ես ոչ եթէ ծանրացայց եւ կամ
տհաճեցայց, այլ ուրախ լել և զինարաւորն ջանացայց տալ ի գործ
եւ յարդիւնս ածել։
թխսիթիւ թօժարուք երթալոյ ի հռոմ, թէ ստուգուբ իցէ. առ
հնարագիտուեն է. զի յերկար ջանացել և զամ ճարտարուք ի գործ
արկել իբրև չկարաց առնուլ զհրաման խոստովանուեն ոպ և կա
մէր. յօժարի արդ երթալ ի հռովմ յուսով թերևս ամբաստանելոյ

Figure 2: Result achieved by Calfa on the microfilm P191 (BnF) - 95.1% good recognition.

Since 2022, Calfa has been in partnership with the Mekhitarist Fathers of Venice for the processing of the correspondences of the Mekhitarist Fathers from 1759 to 1850, in collaboration with Father Vahan Ohanian. The objective is to automatically transcribe several thousand letters in order to speed up the editing and publication work. They feature a wide variety of layouts and are written in cursive, highly abbreviated scripts. For each hand, a new specialized model is produced after transcription of a representative sample. To meet the editorial purposes of the project, models are trained to expand abbreviations, where most errors are located. The models obtained oscillate between 95% and 99% good recognition (see Figure 3). A complete presentation and description of the results will be carried out at the end of the project.



և որդիX և բժ. խնդնէ յացտնի ելել ի վր զօջց մերոց զրզմս ի նցնէ
սպանին։ և յզվա ի նցնէ արրին կալւկանց. և ածին առջի մերոյ
թգրուեX նոյնպս և ուր հանդիպէին փախըXտկն առնէին զնս։
Եւ իսպառ ջնջեցաք զամ ~ն տռապօլիս զայգիսն և զձիբենի
սն. և զբուրաստանս նց առ հսրկ կոտորցք. և աւեր դիձուցք
և արմատախիլ արրք զամ զւռն. և ափրիկեցիքն որ էին անդ
և համարձակեցն և ելին ի պտրզմ ը դէմ մեր։ Յայն ժմ արձակէ
ցաք ի վր նց և կոտորցք XX հսրկ զաՀսն և ապայ առք զմեծ

Figure 3: Results obtained by Calfa on Hurmuz, 1848, Mekhitarist Congregation, 99.01% good recognition.

Conclusion

OCR and HTR technologies are now sufficiently mature to be implemented in the preservation and promotion process of institutions for their collections. The massive digitization of documents, a dynamic in which Armenian institutions such as the Fundamental Scientific Library of NASRA[6], the National Library of Armenia, the Mekhitarists of Vienna (with the support of the Gulbenkian Foundation and the Fundamental Scientific Library) and worldwide institutions participate actively, now provides access to a whole section of Armenian history and literature. Even though imperfect, the results achieved by OCR/HTR technologies are precise enough to allow for their integration into searchable databases, increasing their accessibility and contributing to the sustainability of heritage. Text recognition is not a goal in itself but a step in the creation of a digital heritage and its exploration through digital humanities.

Calfa is a company, based in Paris, and specialized in the automatic processing of Oriental languages (OCR, HTR, text analysis). It develops tailor-made OCR / HTR models and supports heritage institutions in their digitization projects. To find out more about Calfa, the projects carried out and its commitment to heritage: https://calfa.fr.

Acknowledgements

We want to extend our thanks to the Congregation of Mekhitarist Fathers and Father Vahan especially with whom we have been collaborating for several years, as well as to the Fundamental Scientific Library and its scientific advisor Tigran Zargaryan with whom Calfa has a partnership to strengthen OCR models for historical printed documents. The development of OCR was originally supported by the Gulbenkian Foundation in 2016. This communication has been translated by Aliénor Decours-Perez (Calfa).

ԱՄՓՈՓՈՒՄ

Օպտիկական տառաճանաչումը (OCR) և ձեռագիր տեքստի ճանաչումը (HTR) այժմ պատրաստ են գործարկման հայերենի համար։ Այս տեխնոլոգիան կարող է ապահովել փաստաթղթերի ավելի բարձր արժևորում՝ ապահովելով բարելավված հասանելիություն, օգտագործելով, օրինակ, բանալի բառերով որոնում և կարող է թելադրել թվային գրադարանների նոր մարտահրավերներ։

Զեկուցման նպատակն է, ներկայացնելով հայոց լեզվով տեքստերի ճանաչման գործընթացի ժամանակ առաջացած մարտահրավերները, ցույց տալ ժամանակակից հնարավորությունները։ Շեշտադրումն արվելու է ձեռագիր արխիվի, հնագույն ձեռագրերի և հնատիպ գրքերի համար Կալֆայի կողմից մշակված տեխնոլոգիային։

Մենք կներկայացնենք մեր մեկնաբանությունները երեք ընթացիկ նախագծերի՝ Վենետիկի Մխիթարյան միաբանության հայկական ձեռագրերի գրացուցակի, ՀՀ ԳԱԱ հիմնարար գրադարանի թերթերի, Մխիթարյան միաբանության անդամների հայկական նամակների վերամշակման վերաբերյալ։ Կալֆայի կողմից կիրառվող մեթոդաբանությունն ապահովում է ձեռագիր փաստաթղթերի առավել քան 98%-ի և տպագիր փաստաթղթերի առավել քան 99,9%-ի ճշգրտությունը։

REFERENCES

1. Kindt B., Vidal-Gorène C., From Manuscript to Tagged Corpora. An Automated Process for Ancient Armenian or Other Under-Resourced Languages of the Christian East // Armeniaca. International Journal of Armenian Studies, 2022, No 1, pp. 73-96.
2. Kahle P., Colutto S., Hackl G. and Mühlberger G., Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents // 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, 2017, pp. 19-24.
3. Lucas N., Salah C., and Vidal-Gorène C., New Results for the Text Recognition of Arabic Maghribī Manuscripts - Managing an Under-resourced Script // arXiv preprint, 2022, arXiv : 2211.16147.

[6] http://arar.sci.am/

4. Nikolaidou, K., Seuret, M., Mokayed, H. et al., A survey of historical document image datasets // International Journal on Document Analysis and Recognition (IJDAR), Springer, 2022, No 25, pp. 305–338.
5. Ströbel P. B., Clematide S. and Volk. M., How Much Data Do You Need ? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR // Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, ACL Anthology, 2020, pp. 3551-3559.
6. Vidal-Gorène C., Dupin B., Decours-Perez A. and Riccioli T., A Modular and Automated Annotation Platform for Handwritings : Evaluation on Under-Resourced Languages // Document Analysis and Recognition – ICDAR 2021, Cham, Springer, 2021a, pp. 507-522.
7. Vidal-Gorène C., Lucas N., Salah C., Decours-Perez A. and Dupin B., RASAM - A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi // Document Analysis and Recognition – ICDAR 2021 Workshops, Cham, Springer, 2021b, pp. 265-281.
8. Vidal-Gorène C., La reconnaissance automatique d'écriture à l'épreuve des langues peu dotées // Programming Historian, Vol. 5, en français, 2023, [Electronic Publication] URL: https://doi.org/10.46430/phfr0023.