Armenian HTR: State of the art, transcription guidelines and good practices

Chahan Vidal-Gorène, Aliénor Decours-Perez, Anahide Kasparian, Ani Tanelian, Agnès Ohanian

Armenian HTR
State of the art, transcription guidelines and good practices

Chahan Vidal-Gorène[1,2], Aliénor Decours-Perez[1], Anahide Kasparian[1], Ani Tanelian[1], Agnès Ohanian[1]

[1]Calfa, Paris, France,
[2]École nationale des chartes-PSL, Paris, France
Correspondence: contact@calfa.fr

Abstract

This paper discusses the state of the art in Armenian Handwritten Text Recognition (HTR), providing transcription guidelines and good practices. It highlights challenges posed by the Armenian scripts and offers solutions to handle abbreviations and ideograms, and to improve HTR accuracy, focusing on a dataset from the Dulaurier collection.

Introduction

The Armenian language, an independent branche of the Indo-European languages, presents a unique handwriting system that has been in use since the early 5th century. Created by the monk Maštoc' in 405 AD, this alphabet has given rise to a rich tradition of copying, including gospels, liturgical books, hagiographies, historical texts, and legal works. Today, there are around 31,000 manuscripts prior to the 19th century, kept primarily at the Matenadaran in Armenia (11,000), in Vienna (4,000), in Venice (4,000) and in Jerusalem (3,000)[1]. Since the 18th century, numerous handwritten archives originating from the Armenian diaspora, whether for mercantile, intellectual, or religious purposes, have enriched this handwritten heritage, especially in the Ottoman Empire, where printing was forbidden. Armenian is written from left to right and its alphabet consists of 38 letters (36 initially, with 2 additions in the Middle Ages).

Today, the Unicode table for Armenian holds 91 characters: 38 uppercase and lowercase letters, 9 punctuations or intonation marks, 3 religious symbols, 2 phonetic notations, and a ligature (see Table 1). It does not cover all the symbols that can be found in the manuscripts (see infra The Abbreviation System). Over more than 1,500 years of handwritten heritage, the Armenian language has undergone several linguistic evolutions. The so-called 'Classical' Armenian language covers a period from the 5th to the 19th century, when the Modern Armenian was standardized in a Western variant (dialect from Constantinople) and an Eastern variant (dialect from Yerevan). However, Classical Armenian, strictly speaking, encompasses the language used in the gospels, which evolved in the 7th century under the influence of the Hellenizing school (with the creation of numerous abstract neologisms and syntactic constructions mirroring Greek) and further evolved into Middle Armenian in the 11th century within the Armenian Kingdom of Cilicia. Besides the religious texts and some classical works, the manuscripts display many linguistic variations, sometimes combining several forms (for instance, in the clerical correspondence, a mix of Classical Armenian syntax and prepositions with influences of Modern Armenian vocabulary can be observed).

Despite the challenges posed by the uniformity and variations of the Armenian language, Handwritten Text Recognition (HTR) has proven effective and is used in numerous scientific, cultural heritage, and economic projects, achieving very high recognition rates. This success has been achieved

[1]The catalogs, some still incomplete, are available online: for Vienna and Jerusalem see: http://serials.flib.sci.am/matenagitutyun/test/index.html; for the Matenadaran see: https://matenadaran.am/en/matenadaran/digital-resources/grand-catalogue-of-manuscripts/; and for Venice see: https://catalog.mechitar.org.

Letters	Ա ա	Ժ ժ	Ձ ճ	Ր ռ	Օ օ
	Բ բ	Ի ի	Մ մ	Ս ս	Ֆ ֆ
	Գ գ	Լ լ	Յ յ	Վ վ	
	Դ դ	Խ խ	Ն ն	Տ տ	
	Ե ե	Ծ ծ	Շ շ	Ր ր	
	Զ զ	Կ կ	Ո ո	Ց ց	
	Է է	Հ հ	Չ չ	Ւ ւ	
	Ը ը	Ձ ձ	Պ պ	Փ փ	
	Թ թ	Ղ ղ	Ջ ջ	Ք ք	
Intonation marks	՞ (interrogation) ՛ (exclamation) ՝ (emphasis)				
Punctuation marks	: (full point) ՝ (short break) . (long break) , (comma)				
Ligature	և				

Table 1: Letters and common characters in the Armenian unicode system

through specialized projects and a strict set of specifications and requirements for transcription work. This article will outline the specificities and recommendations as guidelines for processing these scripts, relying on the data and models of the *Valorisation Numérique du Fonds Dulaurier project* (BnF Datalab 2023)[2].

1 Digitized corpora and data for HTR

Among the 31,000 Armenian manuscripts (estimation), near 2,000 have been digitized and made available online via IIIF compatible digital libraries, according to the *Index of Digitized Armenian Manuscripts* that has been compiled by Vidal-Gorène, Sargsyan and Van Elverdinghe since 2018 (Vidal-Gorène et al., 2024)[3]. It mostly consists of the collections of Western national libraries (e.g. the Bibliothèque nationale de France with 181 digitized manuscripts or the Biblioteca Apostolica Vaticana with 135 manuscripts), or digitizations made by the Hill Museum and Manuscript Library in partnership with Middle Eastern institutions (for a total of 1,200 manuscripts). Some of these digitization are on microfilms (Bibliothèque nationale de France, BnF) but the re-digitization of the manuscripts with updated standards of quality is ongoing. The proportion of digitized manuscripts that are not available is much bigger: Vidal-Gorène and Decours-Perez (Vidal-Gorène and Decours-Perez, 2020), in their study of the Armenian handwritten heritage in the digital age, published in 2020, were highlighting an estimation of a 5% per year increase of the volume of Armenian digitized manuscripts, with approximately 50% of the Matenadaran collection already digitized to date.

Armenian institutions (e.g. the National Library of Armenia, the Fundamental Scientific Libraric of NASRA, or the Congregation of the Fathers Mekhitarists) are focusing on the digitization of the contemporary archives and the printed press of the 19-20th century, supported by the Calouste Gulbenkian Foundation, involved in the preservation of Modern Western Armenian[4].

Two parallel initiatives for HTR solutions for ancient manuscripts have been developed: starting in 2015 at the University of Vienna with Transkribus (Kahle et al., 2017), focused on historical chronicles, and starting in 2017-2018 with Calfa[5] (partnership with the Musée arménien de France-IRHT, then with a pilot project with the Bibliothèque Universitaire des Langues et Civilisations for the Edouard Dulaurier's collection[6]). Although analysis models for Armenian handwritten documents are available for free online[7], the progress of HTR solutions for Armenian, made possible thanks to the multiple projects of valorization for very specialized collections, is relying on documents that are not in the public domain, hence a lack of data in open access.

Calfa and GREgORI have generated a versatile dataset for Classical Armenian, within the BnF Datalab 2023. This dataset, under pageXML format, consists in 42 images (dig-

[2]Results available at: https://calfa.fr/bnf-datalab-armenian-collection

[3]Available online here: https://www.armenian-manuscripts-index.com

[4]See the Pan-Armenian Digital Library : http://arar.sci.am

[5]https://calfa.fr

[6]Chahan Vidal-Gorène (2 October 2019). Les humanités numériques au service des études arménienne : le partenariat Calfa – BULAC. Le Carreau de la BULAC. Consulted on 3th August 2024 at: https://doi.org/10.58079/m5md

[7]Via the Calfa Vision platform, a web-based annotation tool for documents and images designed for Eastern scripts: https://vision.calfa.fr

itized microfilms) of 14 manuscripts from the Dulaurier collection of the Bibliothèque nationale de France, for a total of 1,467 transcribed lines, following the recommendations described in this article in the section Transcription and annotation guidelines. This dataset is focused on the Armenian medieval historiography, encompassing chronicles from the 8th century (Łevond) to the 13th century (Kirakos Ganjakec'i) and incorporating a sub-sample of Armenian clerical correspondences from Anatolia (19th century). Although mainly based on the hands of Edouard Dulaurier and his students, the dataset is representative of three Armenian scripts (*bolorgir*, *notrgir*, and several variations of *šłagir*, see infra The Armenian Scripts), covers a large set of vocabulary and anthroponyms, and constitute an effective baseline to train HTR specialized models (see infra Efficiency of HTR Models in Armenian).

2 Writing System, Abbreviations and Paleographic Specificities

The Armenian handwritten tradition, while rich and significant, exhibits a relatively limited diversity in scripts and page layouts. This apparent simplicity leads to considerable intra-script variability and substantial textual ambiguity, which can hinder legibility.

The Armenian Scripts Four types of scripts are commonly accepted and used to describe documents, that have been extensively described by Michael Stone et al. in their *Album of Armenian Paleography* (Stone et al., 2002)[8]:

- the *erkat'agir* script, a capital script, often called uncial, used until the 11th century. It is characterized by large letters, few abbreviations, little punctuation, and constant scriptio continua. The letters are separated and distinct without ligature;
- the *bolorgir* script, a very regular bicameral script, that massively introduces punctuation marks and abbreviations. The scriptio continua is less perceptible, even so there is still no clear separation between words and the justification often arbitrarily creates cuts within words. The ligature և (combination of ե + ւ) is generalized and the ligature ﬓ is introduced;
- The *notrgir* script, also known as the notary script, remained in use until the 17th-18th century. This cursive script features smaller lettering and incorporates numerous ligatures and abbreviations, often unique to each copyist. While the letters remain distinct, the extensive use of ligatures for certain combinations of stems and strokes results in a denser textual appearance.
- the *šłagir* script, a cursive and ligatured script that entails all modern and contemporary scripts. This denomination gathers quick-writing scripts often with a very limited ductus and letters linked to each other (beginnings and endings of letters are joined together).

Often, manuscripts are mixing and combining several scripts within a same page (e.g. the *erkat'agir* used strictly for the heading). Influences between scripts can also be observed, like the use of *notrgir* ligatures in manuscripts written in *bolorgir* script. In printing, the *erkat'agir* script is used for uppercase letters, the bolorgir for regular lowercase letters and the *notrgir* for italic script.

The Abbreviation System In comparison with Latin or Middle French, languages that display an impressive volume of abbreviations, each tailored to a specific use or word, the abbreviation system in Armenian seems somewhat perfunctory. The abbreviations are created either through suspension or contraction, and the abbreviated words are only marked either with an horizontal stroke overhead ֊ (named *badiw*), or with a double apostrophe ״ when a vowel is skipped, usually the

[8] Some recent perspectives in computational paleography, conducted by Vidal-Gorène and Decours-Perez since 2021 (Vidal-Gorène and Decours-Perez, 2021) have been presenting the relevance of a seven script-system for the typology and processing of Armenian manuscripts.

There are three abbreviation marks: ˜ can be used for one or several letters, ˝ usually restricted to vowels and ı for the vowel ա / *a*. The transcription is expanding the abbreviated forms with brackets [ ] to include the restored characters. If a character is illegible or if a ligature is preventing to discriminate the letters, these are also introduced within brackets.

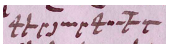
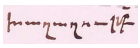
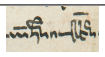
Type	Image	Transcription	Image	Transcription
Suspension		ք[ան]		ամ[ենայն]
Suspension		և [այլն]		Հ[այր].
Contraction		գերյարգու[թ]ե[ան]դ		նմ[ա]ն
Contraction		խաղաղութ[եամ]ք		ը[ստ այն]մ
Ligature		ը[ստ], lig. de ըս		ը[նդ], lig. de ըդ
Ligature		[պետ], lig. de պտ		Պ[ատաս]խ[ան], lig. de Պխ
Ligature		[այսինքն], lig. de այ		ա[մենայն], lig. de մ ˜
Contraction ligature		ա[ւր]հնութե[ան]ն, lig. de հն		թ[են]է, lig. de թէ
Sign ı		իմ[ա]ն[ա]լ		

Table 2: Guidelines for transcription of abbreviations through suspension and contraction, and common ligatures.

vowel ա / *a* (in most cases)[9]. The vowel ա can also be reduced to its beginning stem.

This outward simplicity impairs legibility when the vocabulary or the context is unknown. The use of suspension is limited to some very common prepositions like ք / *k'* for ք-ան / *k'-an*, ը for ը-ստ / *ə-st* or ը-նդ / *ə-nd*, or for the words' endings like substantives ending in -ութիւն / *-ut'iwn*, commonly written -ութի / *-ut'i* in the nominative singular.

The use of contraction varies much more: abbreviated forms are created through the deletion of the vowels of the words, as well as some consonants depending on the needs and wishes of each copyist. It can also results from added ligatures between letters. Originally restricted to nomina sacra[10], the contraction, besides the gospels, can be used for all words, and the proportion of abbreviated words in a text can reach up to 90%. For instance, the word երկրորդ / *erkrord* can be abbreviated եկդ / *ekd*, երկդ / *erkd*, երկրդ / *erkrd* or even երկրրդ / *erkrrd*, that can be easily confused with the words երկ[ի]րդ / *erk[i]rd* or եկ[եալ]դ / *ek[eal]d* (possible from a philologic or contextual point of view, but unlikely form), thus illustrating the high ambiguity of this system. An abbreviated word can also be inflected, leading to apophony and requiring context to be analyzed and understood correctly: the word խորհուրդ / *xorhurd* abbreviated in խհդ / *xhd* will in genitive plural խորհրդոց / *xorhrdoc'* (deletion of the middle ու / *u*) be abbreviated as խհդոց / *xhdoc'*, the deletion of *u* not necessarily explicit and recognized by the HTR model. Table 2 provides keys for reading abbreviations and lists the most common ones.

Ideograms Besides the creation of abbreviations that vary according to their habits of

[9] In rare instances, there has been some abbreviated forms where several letters have been skipped and marked with an apostrophe ˝. See the second to last example in Table 3, where the apostrophe replaces ես / *ēs*. The overhead mark is placed above the deleted letters when there is enough space, but it is not unusual for the abbreviation sign to be misplaced.

[10] For the nomina sacra (God, Lord, Jesus, Christ, Jerusalem, Saint), the contraction is created by only keeping the first and final letters.

copy and the material constraints they are faced with (e.g. lack of space in the page), the Armenian copyists are using ideograms for some common words or words belonging to a specific lexical field. The words երկիր / *erkir* (land, earth), երկինք / *erkink'* (heaven) and աշխարհ / *ašxarh* (world, country) are respectively represented by ⸗ , ⸘ and ⸦ that are the most frequently replaced by ideograms, outside of religious works. Then, as commonly used are the words: արեգական / *aregakn* (sun) represented by ⸙ , լուսին / *lusin* (moon) represented by ⸚ , and աստղ / *astł* (star) represented by ⸛ , with figurative drawings reminding of hieroglyphs. Finally, the adverb որպէս / *orpēs* (like, just as) is usually replace by the symbol ⸜ .

In his work, *Hayoc' gir ev grč'ut'yun* (The Armenian letters and scripts) published in 1973, Abrahamyan counted over 300 ideograms (Abrahamyan, 1973)[11]. Their forms are rarely unequivocal and can vary depending on the copyist. For example, փայլածու / *p'aylacu* (the planet Mercury) can be represented by several distinct ideograms depending on the date of copy: ⸝ , ⸞ , ⸟ , ⸠ and ⸡ . Just like the words they are replacing, the ideograms can be inflected or be used with a definite article, leading to apophony:

- ⸦աց for աշխարհ-աց / *ašxarh-ac'*, genitive plural;
- ⸘ի for երկր-ի / *erkr-i* and not երկիր-ի / *erkir-i*, the first vowel *i* is not accentuated here and is deleted.

When the use of an ideogram appears as less obvious to the copyist, the final letters can be added to the ideogram, even so they are supposed to be included. The following forms are often found:

- ⸦հն for աշխարհ-հն (sic) / *ašxarh-hn*, with a redundant *h* and the definite article *n*;
- ⸦րհաց for աշխարհ-րհաց (sic) / *ašxarh-rhac'*, with a redundant *rh* and the genitive plural desinence.

Monograms The monograms are used, admittedly scarce, but often to write down the anthroponyms or titles in the seals and stamps of modern and contemporary documents. They consist in several letters with ligatures in the same vein as the inscriptions. For instance, the graphical form ⸫ is composed of seven letters and represents the name Ալեքսան / *Alek'san*, or ⸬ with five letters for the name Գէորգ / *Gēorg*. There are instances where ideograms have been created to write down some proper names, ⸭ for Յակոբ / *Yakob* (James) but they constitute an hapax.

Further paleographic considerations The Armenian language has a very elaborate punctuation system, consisting of short breaks, long breaks and a full point. The form and position of these points has varied over time, at times placed overhead (*vernakēt*), at times in the middle (*miǰakēt*), at times on the base line (*kēt* or *storakēt*). There is also a rich system of neumes and of intonation marks, whose use, for some, has been lost. At the ligature level, the sequence of upstrokes and downstrokes or strokes and stems are more likely to be linked in *notrgir* and *šłagir*, for instance ﬔ for թէ / *t'ē*, or ﬓ / ﬗ for մն / *mn*. The letters are generally all linked in the *šłagir* script.

3 Transcription and annotation guidelines

There are no standards for the transcription of the Armenian language, only habits inherited from the edition of texts on the development of abbreviations. In view of the small amount of existing data and of the difficulties to transcribe, the success of HTR for Armenian relies on the limited number of different classes to be recognized.

The increase of symbols is limited by several rules:

- the development of abbreviations,

[11] The manuscript BER Ms. or. quart. 805 of the Staatsbibliothek zu Berlin holds a list of ideograms on folios 271r-273r: https://www.qalamos.net/receive/DE1Book_manuscript_00005146. Some are presenting seldom encountered morphologies and specific to the copyist of this manuscript.

It is recommended to transcribe ideograms and written in full, with the help of brackets to expand the added letters (e.g. *abbr[e]v[iation]*) as for the abbreviations. An ideogram can be replaced by a single symbol, provided that this symbol doesn't cover an apophony due to a grammatical case, or is the opportunity for a legibility mistake with a redundant desinence. If a symbol is used, the brackets should be used and its transcription should be provided distinct with a | (e.g. *[symbol|transcription]*). Some examples with the most common ideograms are presented hereafter.

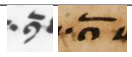
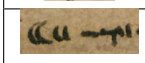
Ideogram	Image	Transcription and remarks
⁓		յ[աշխարհ]իդ ou յ[~|աշխարհ]իդ. There can be erroneous forms, with for example an additional հ besides the ~: to be transcribed strictly by [աշխար]հ, otherwise it could be read as [աշխարհ]հ if transcribed ~հ.
⸓		The nominative and accusative and dative non-erroneous forms can be transcribed [⸓|երկիր].
		To transcribe by յ[երկ]րէ and not յ[երկիր]րէ or յ[երկր]րէ.
ᵽ		The word երկինք, is a plurale tantum, that can be transcribed [ᵽ|երկինք] in nominative plural. The erroneous forms with a redundant ք are to be transcribed [երկին].
		Starting with the accusative, only the present letters are to be transcribed, here յ[երկին]ս.
		To transcribe [երկնի]ց, the desinences added to the ideogram should be kept outside of the brackets [ ].
		The ideogram can be used for derivatives of the word heaven, here [երկնաւ]որ / *[erknaw]or* (céleste).
֏		Transcribe by [որպէս] ou [ø|որպէս].
☽		The nominative and accusative singular non-erroneous forms can be transcribed [լուսին] or [☽|լուսին]. In case of final ն as in ☽ն, we will conclude that this is the definite article and should be transcribed here [լուսին]ն ou [☽|լուսին]ն.
		In all other cases, only the necessary letters are to be transcribed, here զ[լուս]ին.
		The ideogram can be drawn upside down or be used as the basis for the compound words with the word moon or light, e.g. [լու]սաւորն.
☌		The nominative and accusative singular non-erroneous forms can be transcribed [արեգակն] or [☀|արեգակն]. In case of final ն as in the image, we will conclude that this is the definite article and should be transcribed here զ[արեգակն]ն or զ[☀|արեգակն]ն.
		It is common to notice a final double նն. It is an erroneous form that should be transcribed [արեգակ]նն
		In all other cases, only the necessary letters are to be transcribed, here [արեգակ]անն.
		The ideogram can be used as a basis for compound words with the word sun, e.g. [արեգակն]ապ[էս] or [☀|արեգակն]ապ[էս].
m		Transcribe by [ամենայն] or [m|ամենայն]

Table 3: Guidelines for transcription of common ideograms

ideograms and ligatures with the use of brackets to note the letters restored (e.g. վս → վ[ա]ս[ն]), without the use of additional symbol;

- the simplification of the punctuation (limited to three points . : and ՝) and intonation marks (limited to the interrogation mark ՞), that introduce ambiguities and imprecision in their position and forms; as well as the deletion of neums (no equivalent Unicode signs, impairing legibility and with equivocal positions);

- the differentiated processing of other alphabets: multilingual documents are processed with a dedicated HTR model for each alphabet targeted;

- the choice of not transcribing the illu-

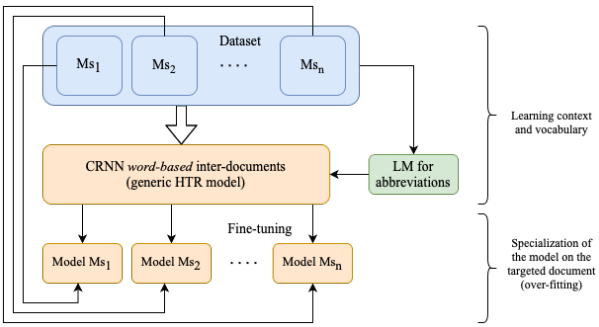


Figure 1: Example of HTR pipeline designed for under-resourced targets and scripts, using a language model to support abbreviations reading, and using a two-steps training strategy

minated letters (drop capitals or not) and diplomatic transcription for numbers with the Armenian alphabet if they are written using it.

The layout of the page and of the text is not dissimilar to the Latin handwritten heritage: notably with the scriptio continua, its use decreased with the gradual introduction of *bolorgir*; the text justification leads to the insertion of spaces within words. The transcription must follow the modern separation between words, thus enabling easy legibility and keywords direct search.

The layout is often limited to one or two columns for the main text, framed by one or several marginalia and a catchword. There are very few glossed documents or with complex layout, apart from tables and schemes. There are illuminated initials and drop capitals, as well as rubrics. The same rules used for the description of Latin documents apply, and SegmOnto ontology (Gabay et al., 2021) is appropriate here.

These recommendations and good practices also apply for incunables and ancient printed documents, notably the ones printed in Venice, Vienna and Constantinople until the beginning of the 20th century, that replicate the layout, the structure and the abbreviation system of manuscripts. A full list of transcription examples is provided in Table 4.

4 Efficiency of HTR Models in Armenian

Vidal-Gorène and Tanelian have presented a state of the art of the HTR performances for Armenian throughout a series of conferences held at the National Library of Armenia in 2023-2024, showcasing recognition rates from 95 to 99% for cursive contemporary archives and ancient manuscripts (Vidal-Gorène and Tanelian, 2024).

To overcome the lack of data and the variability of the abbreviation system in Armenian, a two-step approach has been favored so far: first, a CRNN is trained on words (each word constitutes a class) from a diversified dataset, then, this model is fine-tuned again for each manuscript considered[12] (see Figure 1).

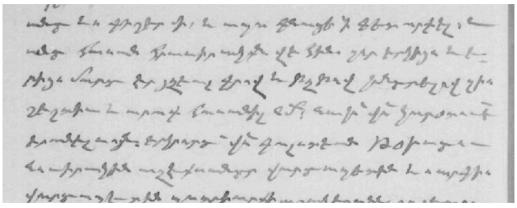
If some cases of overfitting arise, this approach allows for a high rate of recovery for the abbreviations and of good recognition. Applied to the Dulaurier dataset (see supra Digitized corpora and data for HTR), it achieves an average accuracy of 98,56% for the different scripts, including the most cursive and ligatured ones, with a rate of good development for abbreviations of 92,9% (see Figure 2).

The results achieved have been lemmatized and integrated within the search interface of GREgORI[13], with the view to be cross-referenced with other texts and soon directly searchable within Gallica.

Today, the remaining errors from HTR for Armenian are mainly focused on the abbrevi-

[12] This approach at the word level has been validated by Lucas et al. for Arabic Maghribi, that demonstrated a faster convergence of the models in comparison with other state-of-the-art models, including generative-based OCR models (Lucas et al., 2022).

[13] https://www.v2.gregoriproject.com/search/HYE



անդ ևս գրշեր մի, և ապա գնացի ի գետարկել: և
անդ հասան հրաւիրակքն վեհին, զոր երկիցս և ե-
րիցս մարդ էր յղեալ գրով և թղթով խնդրելով զիս
շեշտիւ և արագ հասանիլ նմ: Նախ վ[ե]ր[այ] կարօտու[թ]ե[ան]
երանելւոյներկրորդ՝ վ[ե]ր[այ] գալստեան Թօխադու
նուիրակին աղէքսանդր վարդապետին և սարգիս

Figure 2: Example of HTR model prediction for Armenian on a manuscript from the Dulaurier collection (BnF)

ations, the hallucination of models may lead to often whimsical proposals. The lack of data is also limiting the versatility of the HTR models, without having a detrimental effect on specialized processing. Kindt et al., in their article *An Automated Process for Ancient Armenian or Other Under-Resourced Languages of the Christian East* have demonstrated that for Armenian, a fine-tuning limited to 3 pages was sufficient to achieve a Character Error Rate (CER) of 3% (Kindt and Vidal-Gorène, 2022).

To address the data scarcity that limits the use of larger architectures like Transformers, several strategies have been proposed. One promising approach is data augmentation. Since 2023, Vidal-Gorène et al. have been generating synthetic historical handwritten lines to train HTR models, achieving a 10 percentage point improvement in recognition rates through transfer learning on out-of-domain Armenian manuscripts (similar scripts but different lexical fields) (Vidal-Gorène et al., 2023). This method uses ScrabbleGAN (Fogel et al., 2020), which nevertheless requires a substantial and representative dataset for training. New techniques, such as style mapping, have shown high accuracy on non-historical scripts, and future experiments will explore their application to historical scripts.

5 Conclusion

Although Armenian is an under-resourced language, the challenges of text recognition have largely been overcome, including in few-shot learning situation. The difficulty for HTR for Armenian lies in the simplicity of its paleography: broad-based scripts, high ambiguity for abbreviated text and multiple classes, thus limiting the applicability of HTR models. These specificities are found in manuscripts until the 19th century and more sparsely until the 20th century. However, the guidelines for transcription described in this article allow for the fast convergence of specialized models and an immediate legibility, thus easing the research in big corpora, even for non-specialists. The HTR for Armenian will benefit from massive digitization, collection sharing and data creation following the specifications and requirements described here, and will enable the training of generic and versatile models in the future.

References

Ašot Abrahamyan. 1973. Հայոց գիր եվ գրչության *(= Les lettres et écritures arméniennes)*. EPH, Erevan.

Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. 2020. Scrabblegan: Semi-supervised varying length handwritten text generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4324–4333.

Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, and Claire Jahan. 2021. Segmonto: common vocabulary and practices for analysing the layout of manuscripts (and more). In *1st International Workshop on Computational Paleography (IWCP-ICDAR 2021)*.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, volume 4, pages 19–24. IEEE.

Bastien Kindt and Chahan Vidal-Gorène. 2022. An automated process for ancient armenian or other under-resourced languages of the christian east. *Armeniaca*.

Noëmie Lucas, Clément Salah, and Chahan Vidal-Gorène. 2022. New results for the text recognition of arabic maghribi manuscripts–managing an under-resourced script. *arXiv preprint arXiv:2211.16147*.

Michael E. Stone, Dickran Kouymjian, and Henning J. Lehmann. 2002. *Album of Armenian Paleography*. Aarhus University Press, Aarhus.

Chahan Vidal-Gorène, Jean-Baptiste Camps, and Thibault Clérice. 2023. Synthetic lines from historical manuscripts: an experiment using gan and style transfer. In *International Conference on Image Analysis and Processing*, pages 477–488. Springer.

Chahan Vidal-Gorène and Aliénor Decours-Perez. 2020. Le patrimoine manuscrit arménien à l'ère du numérique : enjeux d'une politique internationale de préservation. In Jean-François Faü, editor, *De la pierre au papier, du papier au numérique*, pages 161–175. Geuthner.

Chahan Vidal-Gorène and Aliénor Decours-Perez. 2021. A computational approach of armenian paleography. In *International Conference on Document Analysis and Recognition*, pages 295–305. Springer.

Chahan Vidal-Gorène and Ani Tanelian. 2024. Գրադարաններ եւ արհեստական բանականութիւն (AI) (= Les bibliothèques et l'intelligence artificielle). In *Libraries and Artificial Intelligence*, Erevan, Armenia. Fundamental Scientific Library of NASRA.

Chahan Vidal-Gorène, Anush Sargsyan, and Emmanuel Van Elverdinghe. 2024. Index of digitized armenian manuscripts.

The following recommendations allow for a reduced number of different classes to be recognized by an HTR model, and to maximize the automatic recognition of Armenian. They enable easy data check / verification and keywords search within digital libraries.

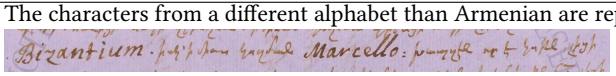
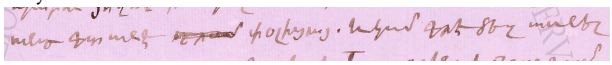
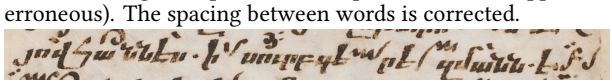
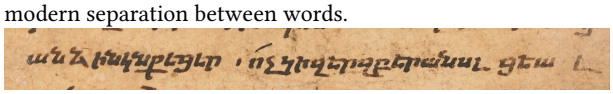
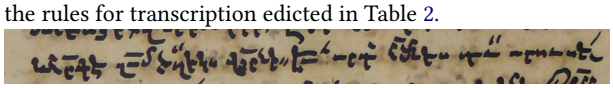
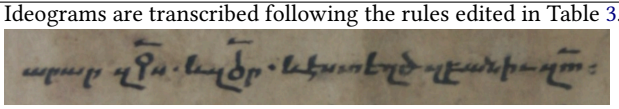
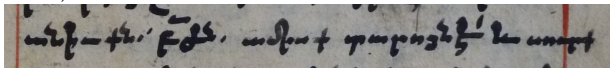
Multi-alphabet	The characters from a different alphabet than Armenian are replaced by #.  Res: #########, իսկ ի միւս կողման ########: Խապզէն որ է կոթն, լիցի
Ratures	The words crossed-out or illegible / unreadable are not transcribed, no matter their position / where they are located.  Res: անդ գտանէ փօլիցայ. և կամ գրէ ձեզ առնել
Intonation marks	The neums, abbreviation marks (” ¯) and intonation marks (´ ˊ) are not transcribed, with the exception of the interrogation point ՞, to be placed where it appears in the text (even if the accentuation is erroneous). The spacing between words is corrected.  Res: յովհա՞ննէս. ի սուրբ գերեզմանն. ե
Ponctuation	The short break (`), the long break or period (.) and the full point (:) are transcribed. The comma (,) and the middle point (•), due to their ambiguity, are transcribed as periods (.). The apostrophe (’) is not taken into account. The hyphen (-) is transcribed as middle-dash (-).
Scriptio continua	Do not follow the scriptio continua or the random spacing used for text justification. Restore the modern separation between words.  Res: անձին կնքեցեր. ոչ կիզեր զբերանս լցեալ
Abbreviations	Abbreviated texts are transcribed through the explanation of abbreviations and ligatures following the rules for transcription edicted in Table 2.  Res: և ն[ո]ր[ո]գէ զամ[ենայն] ծ[ա]ղիկս վ[ա]յր[ե]նի: Եւ արդ՝ նմ[ա]նես դ[ո]ւ արուսե-
Ideograms	Ideograms are transcribed following the rules edited in Table 3.  Res: արար զ[երկինս]. և զ[երկի]ր. և էստեղծ զբանիւ զ[ո|ամենայն]:
Numbers	Numbers are not developed and their notation in digits are kept (keep բժն instead of [երկորտասա]ն).  Res: աիւք. բժն. ամիսք տարոցն է և սուրբ
Initials	Drop capitals are not transcribed.

Table 4: Key recommendations and examples for transcription