e-ISSN 2974-6051

Armeniaca
Vol. 1 – October 2022

From Manuscript to Tagged Corpora
An Automated Process for Ancient Armenian or Other Under-Resourced Languages of the Christian East

Bastien Kindt
Université Catholique de Louvain, Belgique

Chahan Vidal-Gorène
École Nationale des Chartes, Paris

Abstract Creating a digital corpus enriched by full linguistic annotations is a work which classically integrates several manual steps of acquisition, processing, and data display. Processing presupposes the existence of dedicated and specialised analysis tools, adapted to the state of the language used in the corpus. This paper describes a semi-supervised process for building Armenian corpora from scanned documents. This method is based on a chain of applications pre-trained by Calfa and GRE*g*ORI and enabling the complete processing of texts, from their automated input to their linguistic analysis and data display. We provide an assessment of this methodology and benefits of model specialisation, based on digitised copies of a 17th-century manuscript of the Four Gospels (Walters MS W541 = BAL W541, Amida Gospels, ff. 113*v*-117*r*: Lk 1:1-78).

Keywords Handwritten text recognition. Computational philology. Lemmatisation. Morphosyntactic analysis. Tagged corpora. Armenian.

Summary 1 Introduction. – 1.1 Text Recognition. – 1.2 Linguistic Analysis. – 1.3 Aims. – 2 From Handwritten Text to Digitised Text. HTR Processing. – 2.1 Layout Analysis. Identification of Text Area and Line Detection. – 2.2 Text Recognition. – 3 From Digitised Text to Tagged Corpus: Linguistic Analysis. – 3.1 First Step: Analysis by Matching. – 3.2 Second Step: Analysis by RNN. – 4 Conclusion.

Edizioni Ca'Foscari

Peer review

Submitted 2021-12-20
Accepted 2022-03-23
Published 2022-10-28

Open access

© 2022 Kindt, Vidal-Gorène | CC BY 4.0

Citation Kindt, B.; Vidal-Gorène, C. (2022). "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*, 1, 73-96.

DOI 10.30687/arm/2974-6051/2022/01/005

1 Introduction

Online corpora are essential resources for exploring a language and the contents of texts. The current global movement toward digitisation of documents in library collections is leading to the creation of huge image databases. Such initiatives ease access to digitised documents, but they are not sufficient to allow direct and effective researches in the textual data enclosed in these documents. To this end, images must be transformed into texts and texts into tagged corpora. Once they have been enriched with linguistic information and made available through interoperable formats, these corpora can then provide researchers with valuable data and be used for different purposes.

Image databases are increasing in number and growing ever larger. Their conversion into texts and then into corpora must therefore, by necessity, rely on automated processing methods. Two major steps are required to complete such a project: (i) text recognition and (ii) linguistic analysis of corpora. This paper presents and evaluates an operational processing chain developed by Calfa[1] and GRE*g*ORI,[2] and already implemented for different languages of the Christian East.[3] Here, this chain is applied to texts written in Classical Armenian.

1.1 Text Recognition

Good practice in text recognition consists in a three-phase process:

1. layout analysis and understanding;
2. identification of text lines;
3. text extraction itself and its export in a digital format.

State-of-the-art recognition systems achieve excellent results on well-preserved printed documents with simple layouts (Reul et al. 2019). Recognition of historical manuscripts and of complex layouts (e.g. columns, marginal or interlinear scholia, etc.) remains an open problem. The conclusions drawn from the latest *International Conference on Frontiers of Handwriting Recognition* (ICFHR) and *Inter-*

1 Calfa specialises in document analysis for Armenian and other oriental languages; for more information see `https://calfa.fr`.

2 About the GRE*g*ORI project, see `https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html`. GRE*g*ORI has developed an expertise in morpho-syntactic analysis of the main languages of the Christian East, Greek, Armenian, Georgian, and Syriac.

3 Cf. Vidal-Gorène et al. 2020. The GRE*g*ORI Project provides scholars with lemmatised index and concordances, cf. Stone 2021 for Armenian; Schmidt et al. 2021 for Syriac; Pataridze 2020 for Georgian.

national Conference on Document Analysis and Recognition (ICDAR) (Clausner et al. 2019) demonstrate the benefits of using Artificial Intelligence in this field.

Indeed, an artificial neural network can easily be trained with large databases to recognise a given object in a specific context. The success of this approach mainly depends on the availability of large amounts of data, which is not the case for poorly endowed languages, such as Armenian, for which other strategies must be implemented (Vidal-Gorène et al. 2021).

This approach has already proven its efficiency when applied to printed Latin scripts. At present, Optical Character Recognition (OCR) systems have been adapted for Handwritten Text Recognition (HTR) and are being used in a number of digital humanities projects. Figure 1 highlights the common pipeline for the recognition of a handwritten text **[fig. 1]**. For historical manuscripts, layout analysis and character recognition generally achieve 95% or higher accuracy, even with under-resourced scripts and complex layouts (Vidal-Gorène et al. 2021).

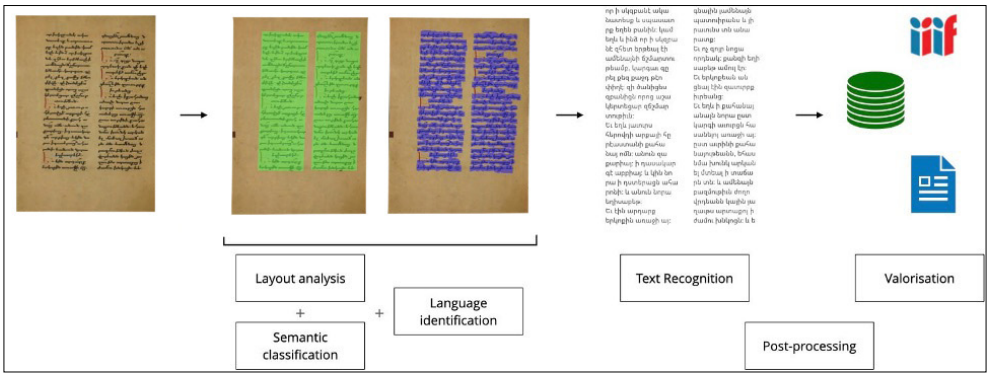


Figure 1 Common steps of HTR process (theoretical aspect): layout analysis, line detection, text extraction and formatting

The recognition engine produces a plain text file retaining the structure of the original document. Such a file already makes it easier to find information (e.g. research by word-form). However, since hyphenation, word spacing, idiosyncratic spellings and mistakes are not resolved, the search possibilities remain limited.

1.2 Linguistic Analysis

Linguistic analysis aims to enrich the textual data with linguistic information. In this case, three types of annotations are carried out:

1. lemmatisation: to assign a lemma (a lexical entry) to each word-form of the text;
2. morphosyntactic tagging: to identify the morphosyntactic category of every word-form (e.g. noun, verb, adjective, pronoun, etc.);
3. inflectional tagging: to provide an analysis for every word-form (e.g. case, number, voice, mood, tense, person, etc.).

For instance, the word զբանիցն (MS W541 f. 113 col. A; Lk 1:4) can be described as follows:

1. զ-: a prefixed preposition (morphosyntactic tag) / զ (lemma);
2. -բանից-: the inflected form of a common noun (morphosyntactic tag) / բան (lemma) genitive (case) plural (number);
3. -ն: the demonstrative suffix (morphosyntactic tag) / ն (lemma).

Tagged corpora make it possible to focus queries on any kind of information recorded within it (word-form, lemma, morphosyntactic and inflectional tags or any combination thereof). This data may then be used for other purposes, paving the way for further studies such as syntactical and semantical analysis.

Such analyses require tools from the Natural Language Processing field (NLP). These tools, first developed for the analysis of modern Western languages in Latin script (written from left to right) are now being adapted to process other languages, including Ancient languages or Oriental languages belonging to different language families or linguistic systems (Indo-European or Semitic languages, inflected or agglutinative languages, etc.) and using different alphabets, or a right-to-left script.

The lemmatisation and tagging steps have initially been implemented with the help of rule-based systems that rely on reference lexicons (built from already analysed corpora) to match the word-forms. This strategy results in effective coverage of the already known vocabulary of the text, irrespective of the context. Resorting to artificial intelligence helps compensate two downsides of this method: lexical ambiguity and unknown terms.

Provided with previously tagged corpora, a neural network (e.g. a Recurrent Neural Network or RNN) can learn, through examples, to infer statistically the analysis of new texts. Instead of manually producing rules for the analysis beforehand, the system is left to generate its own rules using its own devices. The outcomes achieved are therefore predictions. The reliability of the results depends more on the volume and quality of annotations provided initially than on the

sophistication of the rules set out. Using RNN offers a reliable and rapid method for the analysis of new data. This method is particularly appropriate for poorly endowed languages, like Armenian. Despite the existence of several large annotated corpora – Arak29 for Classical Armenian[4] and EANC for Modern Eastern Armenian[5] –, researchers still do not have massive, reliable and interoperable data, as is already the case for modern languages or some ancient languages like Greek.[6]

1.3 Aims

This problematic highlights the need for a full processing chain for text analysis and data creation in Armenian within the scope of the Calfa and GRE*g*ORI projects. These endeavours will significantly increase the extent of computer resources available for the creation of annotated corpora, not only in Armenian, but also in other languages of the Christian East.

In this paper, we describe an experiment conducted on the text of the first chapter of the Armenian version of the Gospel of Luke, as transmitted by the early 17th-century manuscript Baltimore, Walters Art Museum, W541 (= BAL W541), ff. 113*r*-117*r*.[7] The text, in *bolorgir* script, is written with great care. The beginning of the text, on the first folio, partly uses foliate initials and capital letters rubricated in blue and gold (f. 113*r*; cf. **fig. 2**). The text is spread over two columns, containing 23 lines each, with protruding initials. Different intonation and punctuation marks can be seen (cf. **fig. 3**). These initials and interlinear marks may affect text recognition.

4 See `https://www.arak29.am/bible_28E`.

5 See `http://www.eanc.net`.

6 For a state of the art for Armenian, see Vidal-Gorène et al. 2020, 92-5.

7 For a complete description of the manuscript, with high-definition reproductions, see `http://purl.thewalters.org/art/W.541/description`. Images are available under a CC BY 3.0 licence.

Figure 2 MS W541 f. 113*r*

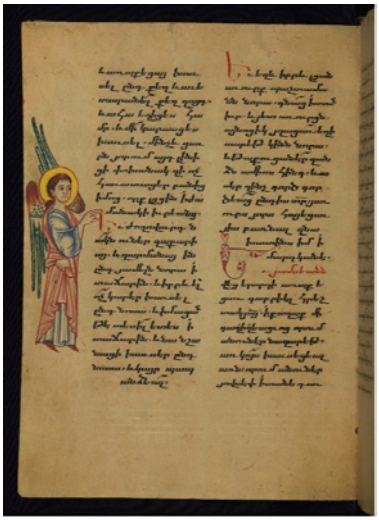


Figure 3 MS W541 f. 114

Regarding the language itself, this text is a sample of Classical Armenian as used in the 5th-century translation of the Gospels. It contains 1,502 word-occurrences and 508 different word-forms, identified after analysis as belonging to 300 different lemmas.

In order to confirm the efficiency of the HTR approach, we also provide results on two other samples 'out of the box' of the same text, with various other difficulties:

1. a page of the 12th-century manuscript W538 (= BAL W538) (ff. 154*r*-156*r*),[8] written in a slanted *erkat'agir*, and following a *scriptio continua* on two columns, sometimes hard to read, and 'text alignment' leading to wrong spaces added into characters of a single word;
2. a page of the printed edition of the Zohrab (1805), from the public domain. The 1805 edition is particularly known for being hard to read due to typography, text density and scan quality.

Experiments are led within the scope of very under-resourced projects, for which we observe a lack of annotated data or a need of specialised transcription.

8 See `https://www.thedigitalwalters.org/Data/WaltersManuscripts/html/W538/`, images available under a CC BY 3.0 licence.

Figure 4a (i) Layout analysis with text-regions identification, (ii) line detection, and (iii) line extraction (MS W541, f. 114*v*). The user keeps control over each step on Calfa Vision in order to ensure high recognition rates

Figure 4b (i) Layout analysis with text-regions identification, (ii) line detection, and (iii) line extraction (MS W538, f. 156*r*)

2 From Handwritten Text to Digitised Text. HTR Processing

The creation of high-performance models for text recognition of ancient manuscripts has not yet been sufficiently evaluated. An effective approach consists in building models specialised on one script or one hand, and then to proceed to fine-tuning, by adjusting the models to the needs of the task at hand (e.g. identification of a text-area in particular, processing an unprecedented abbreviation system, etc.). To be relevant, this methodology should require a dedicated interface to display results and to enable proofreading in order to fine-tune the integrated model **[figs 4a-b]**.

Figure 4c (i) Layout analysis with text-regions identification, (ii) line detection, and (iii) line extraction. Zohrab Bible, 105 (Venice, 1805)

In this case, the document analysis, the semi-automated transcription, and the proofreading of results are undertaken on Calfa Vision,[9] an online semi-automated service specialised in the processing of handwritten documents. The platform allows the creation of customised models for under-resourced languages, for which a massive data approach is limited. Calfa Vision integrates several generic models for layout analysis and HTR (Vidal-Gorène et al. 2021, 513-17). As mentioned above, the document analysis is a three-step process: layout analysis, line detection and text recognition. The processing was

9 See `https://vision.calfa.fr`.

deliberately divided in three steps in order to allow the user to manage the complete process and to customise each feature according to its needs. Figures 4a-c illustrate these three first steps.

2.1 Layout Analysis. Identification of Text Area and Line Detection

Preliminary text area identification is conducted using the method described in Vidal-Gorène et al. (2021, 514). Areas located by this means are categorised on the basis of their content (main body of text, title, marginalia, etc.) and sorted according to the reading order in Armenian. The identification of text area, in blue and red in figure 4.1, reaches 99.64% accuracy.

The engine then proceeds to recognise the lines of text (cf. **fig. 4.1**, steps ii and iii). Across the eight pages of the text, the *precision* (relation between the number of lines correctly predicted and the total number of lines identified) is 89.08% and the *recall* or relevance (relation between the number of lines correctly identified and the total number of lines expected) is 98.53%.[10]

At this stage, the inaccuracies and the mistakes must be corrected manually on Calfa Vision (rectifying the shape of a line, deleting or adding a line, etc.). This operation limits the accumulation of errors throughout the process **[fig. 5]**.

Once the layout has been validated, the extraction of lines is achieved automatically, with the help of a surrounding polygon (in blue in figure 5), and the result can be manually corrected on Calfa Vision. This two-step approach (Diem et al. 2017) allows oblique or curved line localisation.

Thanks to real-time proofreading and evaluation of the models' predictions, the corrected data fed back into the models enable their continuous adjustment to the peculiarities of the corpus. Hence, the quality of predictions increases for the processing of the subsequent images.

10 The difference between the two measurements is due to the high rate of errors obtained on the first folio of the text (113*r*), comprised of illuminated letters and surrounding artworks. For that single folio, the *precision* is 30.77% and the *recall* 99.89%. It means that the four lines of text have been identified, even though they are mixed with a large number of lines detected by mistake. *A contrario*, for the entire text, irrespective of the first folio, the *precision* reaches 96.37% and the *recall* 98.36%. The specific layout of the first folio is the issue here.

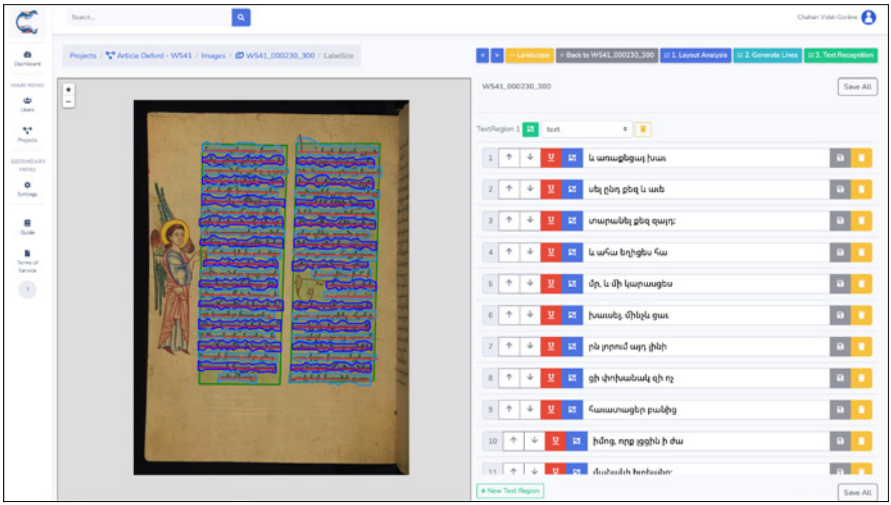


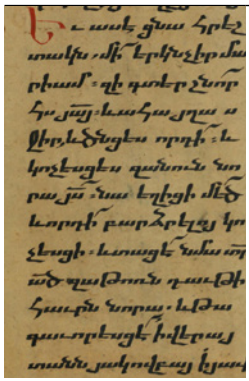
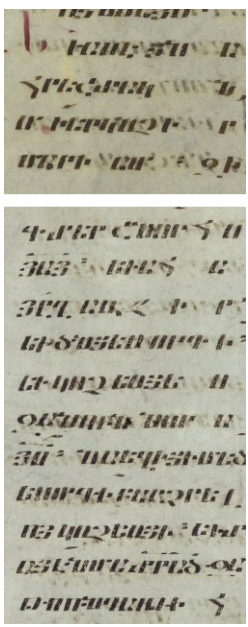
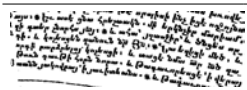
Figure 5 Annotation and proofreading interface, MS W541 f. 114*v*, Calfa Vision (June 2021)

2.2 Text Recognition

Already at the previous step, the identified lines of text can be submitted to the HTR. Calfa Vision includes several generic models of text recognition for the four main types of Armenian handwriting, namely the *erkat'agir*, *bolorgir*, *nōtrgir* and *šłagir* scripts (Stone et al. 2002). The HTR error rate is assessed by a specific metric, the Character Error Rate (CER). The *bolorgir* model used here by default gives a 5.42% CER for the manuscript. Figure 6 shows the confusion matrix displaying the distribution of errors **[fig. 6]**.

Warm colour indicates that a predicted character (on the X-axis) is often transcribed as an expected character (on the Y-axis). For instance, the character ա is well recognised by the HTR, as the cell is red in the matrix. For this character, the outcome is close to 100%. The matrix shows the distribution of the recognition rate for each character (high on the diagonal). It means that a significant proportion of letters is correctly recognised. Different hues indicate characters with lower recognition rates. Such is the case for the letters ք, լ, and յ, whose recognition rate is 70%.

Figure 7 Example of prediction from model by default, and after fine-tuning with additions of spaces and resolution of abbreviations, MS W541 f. 115*r* col. A, MS W538 f. 156*r* col. B, 156 col. A., and the Zohrab Bible (1805), p. 115

Original picture	Default model prediction	Fine-tuned model prediction (after 3 images)
	ւասէգնահրեշ տակն,միերկնչիրմա րիամ:զիգտերշնոր հայայ:ահաւյղաս ջիրևծնցեսորդի:և կոչեսցեսզանուննո րայս:նաեղիցիմեծ: ևորդիբարձրելոյկո չեսզի:ևտացէնմատր ածոյաթոռնդաւթի հաւրննորա.և թա գաւորեսցէիվերայ տաննյակովբայիյաւի տեանս:և թագաւոր	Եւ ասէ ցնա հրեշ տակն, մի երկնչիր: մա րիամ: զի գտեր շնոր հս յաստուծոյ: և ահա յղաս ջիրև ծնցես որդի: և կոչեսցես զանուն նո րա յիսուս: նա եղիցի մեծ: և որդի բարձրելոյ կո չեսցի: և տացէ նմա տր աստուած զաթոռն դաւթի հաւրն նորա: և թա գաւորեսցէ իվերայ տանն յակովբայ ի յաւի տեանս:և թագաւոր
	ԵՒԱՍԷՑՆ Ա ՀՐԵՇՏԱԿ Ն, Մ ԻԵՐԿՆՉԻ Ր ՄԱՐԻ ԱՄ : ԶԻ ԳՏԵՐ ՇՆՈՐՀՍ ՅԱՅ: ԵՒ ԱՀ Ա ՅՂՍՈՒՀ Ի Ր, ԵՒ ԾՆՑԵՍՈՐԴ Ի: ԵՒ ԿՈՉԵՍՑԵ Ս ԶԱՆՈՒՆ ՆՈՐ Ա ՅՍ: ՆԱԵՂԻՑԻՄԵՑ ԵՒՈՐԴԻԲԱՐՁՐԵԼ ՈՑԿՈՉԵՍՑԻ: ԵՒՏ ԱՑԷՆՄԱՏՐԱԾԶԱ ԹՈՌՆԴԱՒԹԻ Հ ԱՒՐ ՆՈՐԱ: ԵՒԹԱ ԳԱՒՈՐԵՍՑԷ	եւ ասէ ցն ա հրեշտակ ն, մ ի երկնչիր մարիամ : զի գտեր շնորհս յաստուծոյ: եւ ահ ա յղասջ ի ր, եւ ծնցես որդի: եւ կոչեսցե ս զանուն նորա յիսուս: նա եղից ի մեծ եւ որդի բարձրել ոյ կոչեսցի: եւտ ացէ նմա տէրաստուած զա թոռն դաւթի ի աւր նորա: եւ թա գաւորեսցէ
	Եւ ասէ ցնա հրեշաակն. մի երկնչիր մարիամ, զի գտեր շնորհս յայ: և ահա յղասջիր և ծնցես որ դի. և կոչեսցեն զանուն նր ՅՍ: Նա եղիցի մեծ, և որդի բարձրելոյ կոչեսցի: և տացէ նմա ար ած զա թոռն դաւթի հօրն նորա. և թագաւոր	Եւ ասէ ցնա հրեշտակն. մի երկնչիր մարիամ, զի դտեր շնորհս յաստուծոյ: և ահա յղասջիր և ծնցես որ դի. և կոչեսցեն զանուն նորա Յիսուս: Նա եղիցի մեծ, և որդի բարձրելոյ կոչեսցի: և տացէ նմա տէր աստուած զա թոռն դաւթի հօրն նորա. և թագաւոր

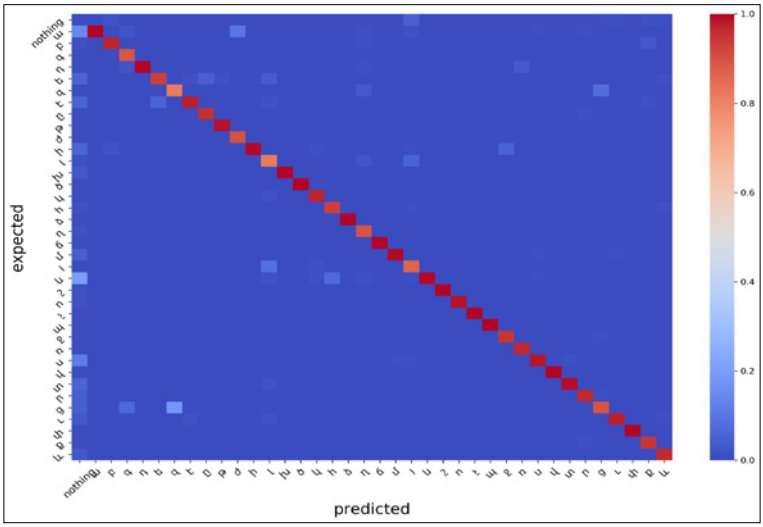


Figure 6 Distribution of HTR errors in the confusion matrix (default model)

With a CER of 5.42%, an overall understanding of the text can be achieved (cf. column Default model prediction in **fig. 7**). Errors are recurrent and located on a limited number of letters (cf. figs 6-7). However, a good character recognition model does not mean that predicted output is directly exploitable as is by researchers, because the text produced by the HTR is limited in its inter-word spaces recognition and it preserves a *scriptio continua* (model originally trained on texts without word spaces), end-of-line word breaks and abbreviations of the original text. Several approaches are possible: on the basis of the obtained text (see fig. 7), either automatically generate word spacing and resolve abbreviations in post-processing (Camps et al. 2021), or manually add the spaces, corrections and desired information in order to fine-tune the models with this new text as a reference. We favour the second approach, because it gives the user total control over its editorial choices, directly on Calfa Vision. Figure 8 shows how this fine-tuning helps reduce the CER, depending on the number of images that undergo manual correction **[fig. 8]**.

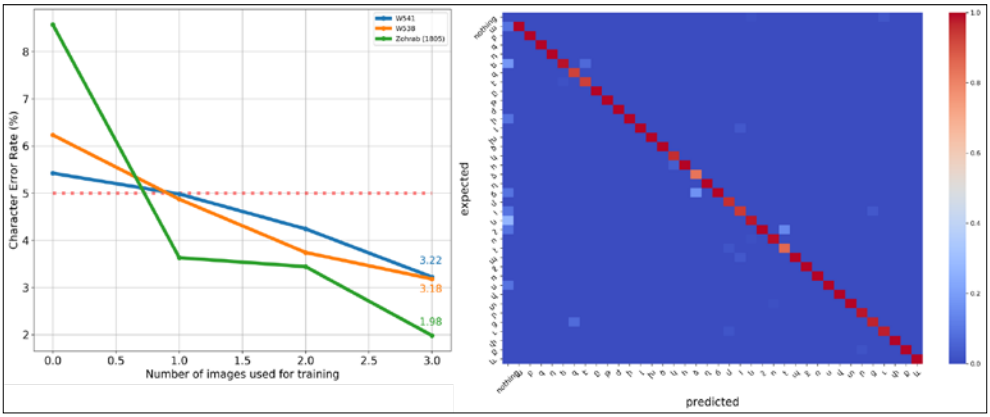


Figure 8 Distribution of HTR errors in the confusion matrix (Fine-tuned model) and CER evolution

The new confusion matrix obtained after correcting three images shows limited information loss for the letters ձ and ղ, as well as a recognition loss for the letter ճ. Nevertheless, the fine-tuning did result in limiting the confusion between characters. The CER is now 3.22% and word separation is 95.42% accurate.[11] This step shows the interest of an automated annotation platform such as Calfa Vision for a customised specialisation, with only three images to manually proofread. It corresponds to 100 very short lines for manuscript MS W541 (only 4 words by line), when state-of-the-art models and technologies generally requires between 600 and 2,500 lines. The same applies for MS W538 and the Zohrab bible.

The text achieved at this stage can either follow a diplomatic transcription or be adjusted to the needs of the user, depending on the choices made during the proofreading of predictions and the potential fine-tuning. However, the text is not standardised, end-of-line word breaks, most notably, being retained in cases where the break is not obvious (lack of a hyphen).

11 We notice a very significant gain in accuracy after a fine-tuning conducted on five to ten corrected images. In this scenario, the CER is below 2.5%. The architecture proposed by Calfa seems efficient to resolve directly various abbreviations at the HTR stage (Camps et al. 2021), with a larger number of images to correct however, not only with three images as we did for W541. For example, the user could decide to transcribe all instances of այ in աստուծոյ or in աստուածոյ or in Ա(ստուծո)յ according to their own editorial choices, and to train the models to replicate this transcription rule. The same applies for other abbreviated words with an abbreviated mark.

3 From Digitised Text to Tagged Corpus: Linguistic Analysis

The digitised texts undergo linguistic analysis, as a result of which each word-form is lemmatised and morphosyntactic features as well as inflectional analysis are provided. To this end, a mixed method was applied: an analysis by matching (using GRE*g*ORI's lexical data), assisted by an analysis by RNN. Outcomes have been compared to the analyses provided by Arak29.

3.1 First Step: Analysis by Matching

The analysis by matching works by comparing the vocabulary of a given text with the lexical data already gathered in reference lists, here the linguistic resources of the GRE*g*ORI project (as described, in Greek, in Kindt 2021, 175-83). For the Armenian language, these resources consist of digital dictionaries of both simple forms and polylexical ones, i.e. with prepositional prefixes and determinative suffixes. In these resources, word-forms are linked with their lexical analysis (lemma), morphological analysis (morphosyntactic category) and inflectional analysis (case, number, voice, mood, tense, person, etc.). They include word-forms attested in the corpora processed earlier, currently amounting to more than 67,039 word-occurrences, 25,000 unique, either simple or polylexical (cf. Coulie et al. 2022). **Table 1** presents a sample of simple word-forms of the lemma ագաւն.

Table 1 Sample of simple word-forms of the lemma ագաւն

Word-form	Lemma	Morphosyntactic analysis	Inflectional Analysis
ագաւն	ագաւն	A	:As:Ns*
ագաւնաց	ագաւն	A	:Âp:Dp:Gp
ագաւնաւ	ագաւն	A	:Hs
ագաւնաւք	ագաւն	A	:Hp
ագաւնէ	ագաւն	A	:Âs
ագաւնի	ագաւն	A	:Ds:Gs:Us
ագաւնս	ագաւն	A	:Ap:Up
ագաւնք	ագաւն	A	:Np
ագաւնօք	ագաւն	A	:Hp

* The GRE*g*ORI Project uses a specific inflectional tagset described in Coulie et al. 2021 ; e.g. "As" = acc. sing., "Ns" = nom. sing., "Âs" = abl. sing., "Hp" = instr. plur., "Up" = locatif plur., etc.

These resources also comprise automatically generated word-forms, in order to complete the inflectional paradigms of some lemmas and fill in standard combinations of simple word-forms with prepositional prefixes and determinative suffixes. All these data, totalling more than 850,000 different word-forms (simple or polylexical), can be considered as a potential lexicon, increasing the lexical coverage during the lexical look-up process. **Table 2** presents a sample of automatically generated word-forms of the lemma ազատ.

Table 2 Sample of generated word-forms for the lemma ազատ

Word-form	Lemma	Morphosyntactic analysis	Inflectional analysis
ազատս	ազատ	A*	:Ap:Up
ազատսդ	ազատ@դ	A@PRO+Dem	:Ap:Up@ø
ազատսն	ազատ@ն	A@PRO+Dem	:Ap:Up@ø
ազատսս	ազատ@ս	A@PRO+Dem	:Ap:Up@ø
ազատսստ**	ազատ@դ	A@PRO+Dem	:Ap:Up@ø
զազատս	զ@ազատ	I+Prep@A	ø@:Ap
զազատսդ	զ@ազատ@դ	I+Prep@A@PRO+Dem	ø@:Ap@ø
զազատսն	զ@ազատ@ն	I+Prep@A@PRO+Dem	ø@:Ap@ø
զազատսս	զ@ազատ@ս	I+Prep@A@PRO+Dem	ø @:Ap@ø
զազատստ	զ@ազատ@դ	I+Prep@A@PRO+Dem	ø@:Ap@ø
յազատս	ի@ազատ	I+Prep@A	ø@:Ap:Up
յազատսդ	ի@ազատ@դ	I+Prep@A@PRO+Dem	ø@:Ap:Up@ø
յազատսն	ի@ազատ@ն	I+Prep@A@PRO+Dem	ø@:Ap:Up@ø
յազատսս	ի@ազատ@ս	I+Prep@A@PRO+Dem	ø@:Ap:Up@ø
յազատստ	ի@ազատ@դ	I+Prep@A@PRO+Dem	ø@:Ap:Up@ø

* The GRE*g*ORI Project uses a specific morphosyntactic tagset described in Coulie et al. 2021; e.g. “N+Com” = (common)-noun, “A” = adjective, “V” = verb, “I+Prep” = prepostion (“I”, for “Invariable”, characterises uninflected words), “PRO+Dem” = demonstrative pronoun or suffix, etc.

** Automated word generation includes uncommon (or inaccurate) spellings attested in manuscripts or editions (in this case unexpected ազատստ instead of ազատսդ); samples in Stone 2021, 21, 93.

At the end of the analysis by matching, the simple word-form եղեն (f. 113 col. A; Lk 1:2) is analysed as եղանիմ.V:MÎJ3p – lemma: եղանիմ; category: verb; morphological analysis: MÎJ3p = mediopassive aorist indicative, third person plural. The polylexical word-form ցիրեշտակն (f. 114*r* col. V; Lk 1:28), which can be split into ց-հրեշտակ-ն, is analysed as ց.I+Prep - հրեշտակ.N+Com:As - ն.PRO+Dem; distinguishing the prepositional prefix (ց-), the noun

(հրեշտակ) in the accusative singular (As), and lastly the determinative suffix (-ն).[12]

This approach of analysis by matching quickly provides very reliable results for the word-forms already compiled in the resources, as well as for non-ambiguous word-forms. In Greek, by using the lexical resources of the GRE*g*ORI Project, it yields a coverage of more than 90% of the vocabulary of a new text (Kindt, Pirard 2016). However, the approach also has limitations that influence its outcomes.

First, if a given word-form is not already listed in the reference resources, no analysis can be provided. The word յաւժարեցին (f.113*r* col. AB; Lk 1:1), missing from the resources, even under the alternate spelling յօժարեցին (lemma յաւժարեմ), has no match. In the first chapter of Luke, the situation mainly concerns proper nouns (anthroponyms or toponyms), such as Եղիսաբեթ (f. 113 col. A; Lk 1:5), Նազարէթ (f. 114 col. B; Lk 1:26), or even abbreviations like իղի for Իսրայեղի (f. 117*r* col. B; Lk 1:80).

Second, the analysis by matching does not take into account the contexts in which words appear. Hence, when several analyses are possible for a single word-form, all of them are returned. For instance, the word կամ (f. 113 col. A; Lk 1:3) is analysed as կամ.V:EÎP1s (verb lemma), կամ (եւ).I+Conj (lemma of the conjunction ‘or’) and կամ (կամաց).N+Com:As:Ns (noun lemma).[13] Some simple word-forms are homographs of word-forms with the demonstrative suffix -ս. The word աւուրս (f. 117*r* col. A; Lk 1:75), for instance, is analysed both as աւր.N+Com:Ap:Up (simple word-form) and as աւուր,աւր.N+Com:Ds:Gs:Us - ս,.PRO+Dem (polylexical word-form).

Last, if not all possible analyses of a given word-form are recorded in the reference resources, the analyses provided this way remain partial and can be erroneous. The words ած (for instance f. 114*r* col. B; Lk 1:16) and այ (for instance f. 117*r* col. B; Lk 1:78), which are actually abbreviated forms of Աստուած and Աստուծոյ respectively, have been analysed as ած,ածեմ.V:EÎJ3s and այ,ահ.I+Intj. Though technically correct out of context, these outcomes are erroneous in this particular case.

The text of the first chapter of the Gospel of Luke in the W541 manuscript is made up of 1,052 words-occurrences. As shown in **table 3**, the resources of the GRE*g*ORI project identified 973 word-oc-

12 At this stage, we can notice that the Arak29 analysis has the following outcome for the polylexical form ցհրեշտակն: հրեշտակ - noun.acc.sg, which only identifies the noun lemma, without acknowledging the prepositional prefixes and determinative suffixes. This linguistic description, more concise than the one provided by GRE*g*ORI, limits the automated comparison with the Arak29 tagging, and hence the experiment.

13 Moreover, in this last case, there are also two possible inflectional analyses: nominative singular (Ns) or accusative singular (As). In this paper, we focus on lexical analyses, leaving aside inflectional ambiguities.

currences, among which 79 word-occurrences are left with no analysis and 211 words-occurrences are assigned to more than one lemma.

Table 3 Outcomes of the linguistic analysis by matching

	Total	Proportion	Examples
Word-forms	1.052	100%	
Analysis = 0	79	7,51%	
Analysis = 1	762	72,43%	
Analyses ≧ 1	973	92,49%	
Analyses > 1	211	20,05%	
Analyses = 2	116	11,03%	պատասխանի = պատասխանի.N+Com:As:Ns vs պատասխանեմ.V:MÎP3s
Analyses = 3	26	2,47%	նմանէ = նա (նա).PRO+Dem:Âs vs նմանն.A:Âs vs նմանեմ.V:EÎP3s
Analyses = 4	60	5,70%	ի = ի.I+Prep vs ինի.N+Lettre vs 20.NUMA+Car vs 20th.NUMA+Ord
Analyses = 5	0	0,00%	
Analyses = 6	9	0,86%	է = է.I+Intj vs եմ.V:EÎP3s:MÎP3s vs է (ա).N+Lettre vs է (էից).N+Com:As:Ns vs է,7.NUMA+Car vs է,7th.NUMA+Ord

Thus, although the analysis by matching covers 92.49% of the vocabulary of the text processed, this result has to be qualified considering the limitations outlined above. These words without analysis and words with more than one analysis should be checked before delivering the final data. This verification step can be executed manually. It is, however, a very tedious and time-consuming task, when done over massive corpora. The analysis by RNN makes it possible to overcome these difficulties.

3.2 Second Step: Analysis by RNN

The RNN model used is the one built by Calfa in March 2020 (Vidal-Gorène, Kindt 2020), using the Pie architecture (Manjavacas et al. 2019). It has been trained with a corpus of 67,039 analysed word-forms from the GRE*g*ORI resources (Coulie et al. 2021). General accuracy of this model is 90.44% for the lemmatisation task (86.20% for the ambiguous tokens and 68.64% for the unknown tokens of the testing set) and 92.39% for the morphosyntactic annotation task (91.45% for the ambiguous tokens and 74.41% for the unknown tokens). This model provides a single prediction for each word-form, including unknown word-forms and word-forms that could have sev-

eral analyses with an approach by matching. **Table 4** displays the outcomes of the analysis by matching and of the RNN predictions for the same sample of text.

Table 4 Sample from the analyses by matching and by RNN (fol. 114 col. A; Lk 1:19-21)

Word-forms	**Analysis by Matching**		**Analysis by RNN**	
	Lemma	**Morphosyntactic Annotation**	**Lemma**	**Morphosyntactic Annotation**
և	և	I+Conj	և	I+Conj
առաքեցայ	առաքեմ	V	առաքեմ	V
խաւսել	խաւսեմ	V	խօսիմ	V
ընդ	ընդ	I+Prep	ընդ	I+Prep
քեզ	դու	PRO+Per2s	դու	PRO+Per2s
և	և	I+Conj	և	I+Conj
աւետարանել	աւետարանեմ	V	աւետարանեմ	V
քեզ	դու	PRO+Per2s	դու	PRO+Per2s
զայդ:	զ@այդ	I+Prep@PRO+Dem	զ@այդ	I+Prep@PRO+Dem
և	և	I+Conj	և	I+Conj
ահա	ահա	I+Intj	ահա	I+Intj
եղիցես			եղանիմ	V
համր,			համր	N+Com
և	և	I+Conj	և	I+Conj
մի	մի(միոջ)	NUM+Car	մի (ոչ)	NUM+Car
կարասցես			կարեմ	V
խաւսել,	խաւսեմ	V	խօսիմ	V
մինչև	մինչև	I+Conj	մինչև	I+Conj
ցաւրն	ց@աւր@ն	I+Prep@N+Com@PRO+Dem	ց@օր@ն	N+Com@PRO+Dem@ø
յորում	ի@ո՞ր	I+Prep@PRO+Int	ի@որ	I+Prep@PRO+Rel
այդ	այդ	PRO+Dem	այդ	PRO+Dem
լինիցի	լինեմ	V	լինիմ	V
փոխանակ	փոխանակ	I+Adv	փոխանակ	N+Com
զի	զ@ի	I+Prep@I+Prep	զի	I+Conj
ոչ	ոչ	I+Neg	ոչ	I+Neg
հաւատացեր	հաւատամ	V	հաւատամ	V
բանիցդ	բան	N+Com	բան	N+Com
իմոց,	իմ	PRO+Pos1s	իմ	PRO+Pos1s
որք	ո՞ր	PRO+Int	որ	PRO+Rel
լցցին	լնում	V	լնում	V
ի	ի	I+Prep	ի	I+Prep
ժամանակի	ժամանակ	N+Com	ժամանակ	N+Com

Results were automatically evaluated by comparing the analyses produced by the two approaches with those provided by Arak29. The RNN approach highlights several points (see **table 5**). We first notice that the RNN fixes 47 analyses produced by the matching step (6.16% error). These are often imprecise or erroneous analyses present in the lexical resources used.

Then, we can observe that the analyses of the RNN are correct in 89.87% of cases for the lemmatisation of word-forms with only one possible analysis, and in 90.52% of cases for the ambiguous word-forms. Regarding the morphosyntactic annotation, these rates reach 60.75% and 92.41% respectively. RNN is therefore more efficient for word-forms disambiguation than for unknown word-forms prediction.

Table 5 Evaluation of the RNN approach

	GRE*g*ORI quantitative data		**GRE*g*ORI wrong analyses**		**RNN correct lemma**		**RNNcorrect morphosyntactic annotation**	
	Total	%	Total	%	Total	%	Total	%
Word-forms	1.052							
Analysis = 0	79	7,51%			71	89,87%	48	60,75%
Analysis = 1	762	72,43%	47	6,16%				
Analyses > 1	211	20,05%			191	90,52%	195	92,41%

The texts of the RNN training set mainly consist in texts with a specific state of Armenian language, the so-called Hellenising school, significantly different from the Classical Armenian of the Gospels.[14] Despite this bias, the model demonstrates a good capacity of generalisation.

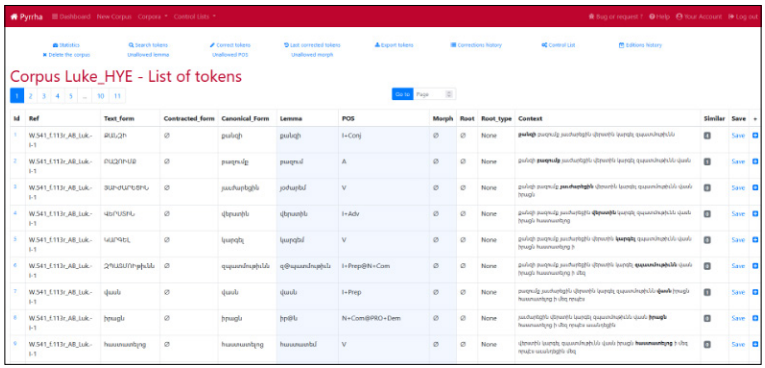
In case of unanalysed word-forms during the matching steps – meaning the word-form is unknown in the GRE*g*ORI resources – the model notably fails on the morphosyntactic analysis of proper nouns (36% of mistakes). For instance, the model analyses erroneously the word Թէոփիլէ (f. 113 col. A; Lk 1:3) as a verb, and the word Զաքարիայ (f. 114*r* col. B; Lk 1:18) as a preposition (զ-) followed by an anthroponym.

However, the model manages without too much difficulty various spelling variations, such as the alternations աւ/օ or լ/ղ. The two words զաբղիէղ (f.114*r* col. B; Lk 1:19) and զաբղիէլ (f.114 col. B; Lk 1:26), missing in the GRE*g*ORI resources, are correctly analysed and lemmatised.

In the end, the defined hybrid approach achieves a correct lemmatisation at 93.06% and a correct morphosyntactic analysis at 91.44%.

14 About this question, see Coulie 1995; Muradyan 2012; Meyer 2018.

These results constitute a baseline for analysis of Classical Armenian and demonstrate the relevance of a hybrid approach when processing an under-resourced language or with very specific data. Two areas of improvement emerge from these experiments: 1. to integrate additional analysed data in the training set, from other texts processed by GRE*g*ORI or from other projects, such as literary quotes from NBHL; 2. it would also be useful, as we did for the text recognition step, to specialise the model on a part of the corpus being analysed and to carry out an analysis by iteration.



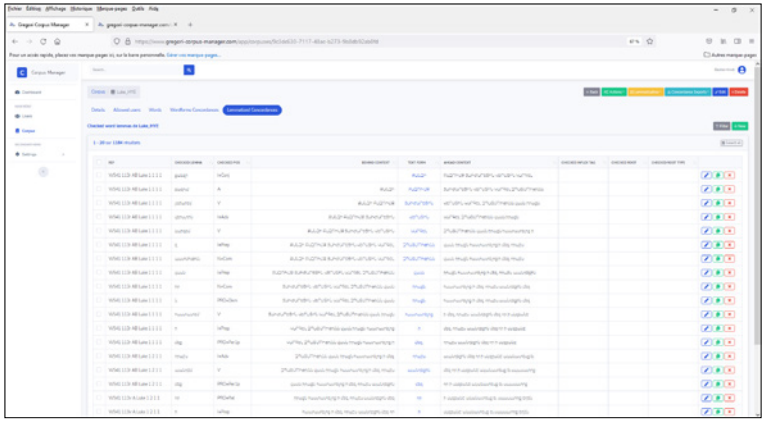


Figure 9 Pyrrha interface

Figure 10 GCM interface

There are tools for visualising and proofreading analysed corpora, such as Pyrrha (Clérice et al. 2019) (cf. **fig. 9**), or the GRE*g*ORI Corpus Manager (GCM) (cf. **fig. 10**), two online interfaces enabling collaborative work. Pyrrha is an interface initially developed by the

École Nationale des Chartes (Paris). The user can view the processed texts and make corrections to the linguistic analyses of the corpus **[figs 9-10]**.

Pyrrha and the GCM share common features. However, GCM can process larger corpora, import lemmatised data produced by other analysis tools and instantly generate word-form concordances or lemmatised concordances.

Pyrrha and the GCM are useful tools to correct analysed data of corpora, partially or in full. These corrections make it possible to enrich the resources used for an analysis by matching or to specialise neural models.

4 Conclusion

In this paper, we evaluate the use of Calfa and GRE*g*ORI tools for the semi-automatised creation of corpora from digitised documents. The strategy of using generic models gradually specialised on the considered task quickly results in a CER of 3.22%, a lemmatisation of 93.06% and POS-tagging of 91.44%.

If these results really depend on the choice of the manuscript and the language state of the processed text, they nevertheless demonstrate the relevance of such an approach with extremely insufficient data (only three images in training for recognition by HTR, an initial corpus of 67,039 word-occurrences for lemmatisation and morpho-syntactic analysis), which is the case for most under-resourced languages or with non-Latin scripts. Evaluations carried out on W538 and the Zohrab bible, as control samples, highlight the adequacy of the process applied to new documents with other kind of difficulties.

The interfaces used for this paper provide interoperable data with other systems and enable full control of the pipeline and of editorial choices. The continuous improvement of generic models is at the heart of the implemented strategy, in order to strengthen the ability of fast specialisation of tools and models. The described processing chain demonstrates the effective capacity of systems implemented by Calfa and GRE*g*ORI to produce corpora and linguistic data, opening new perspectives for under-resourced languages, in general, but also specifically for Armenian studies.

Bibliography

Camps, J.-B.; Vidal-Gorène, C.; Vernet, M. (2021). "Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches". Barney Smith, E.H.; Pal, U. (eds), "Document Analysis and Recognition – ICDAR 2021 Workshops. ICDAR 2021". *Lecture Notes in Computer Science*, 12917, 306-16. https://doi.org/10.1007/978-3-030-86159-9_21.

Clausner, C.; Antonacopoulos, A.; Pletschacher, S. (2019). "ICDAR2019 Competition on Recognition of Documents with Complex Layouts – RDCL2019". *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1521-6. https://doi.org/10.1109/ICDAR.2019.00245.

Clérice, T.; Pilla, J.; Camps, J.-B.; Jolivet, V.; Pinche, A. (2019). "Pyrrha, A Language Independent Post Correction App for POS and Lemmatization". *Zenodo*. https://doi.org/10.5281/zenodo.2325427.

Coulie, B. (1994). "Style et traduction. Réflexions sur les versions arméniennes des textes grecs". *RÉArm*, 25, 43-62. https://doi.org/10.2143/REA.25.0.2003773.

Coulie, B.; Kindt, B.; Kepeklian, G.; Van Elverdinghe, E. (2022). "Étiquettes morphosyntaxiques et flexionnelles pour le traitement automatique de l'arménien ancien". *Le Muséon*, 135(1-2), 207-39.

Diem, M.; Kleber, F.; Fiel, S.; Grüning, T.; Gatos, B. (2017). "cBAD: ICDAR2017 competition on baseline detection". *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, 1355-60. https://doi.org/10.1109/ICDAR.2017.222.

Kindt, B. (2021). "Du texte à l'index. L'étiquetage lexical du De Septem Orbis Spectaculis de Philon le Paradoxographe: méthode et finalité". Labarre, G. (éd), *Sources, Histoire et Éditions. Les outils de la recherche. Formation et recherche en science de l'Antiquité*. Besançon: Presses universitaires de Franche-Comté, 167-210.

Kindt, B.; Pirard, M. (2016). "De Nazianze à Ninive. La couverture lexicale du Dictionnaire Automatique Grec". Somers, V.; Yannopoulos, P. (eds), *Philokappadox. In memoriam Justin Mossay*. Louvain; Paris; Bristol CT: Peeters, 49-77. Orientalia Lovaniensia Analecta. Bibliothèque de Byzantion 25 | 14.

Manjavacas, E.; Kádár, Á.; Kestemont, M. (2019). "Improving Lemmatization of Non-Standard Languages with Joint Learning". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 1493-503. http://dx.doi.org/10.18653/v1/N19-1153.

Meyer, R. (2018). "Syntactical Peculiarities of Relative Clauses in the Armenian New Testament". *RÉArm*, 38, 35-83. https://doi.org/10.2143/REA.38.0.3285778.

Muradyan, G. (2012). *Grecisms in Ancient Armenian*. Leuven: Peeters Publishers. Hebrew University Armenian Studies 13.

Pataridze, T. (2020). *Vie et conduite des Bienheureux Justes-nus et de notre saint Père Zosime: trois traductions géorgiennes*. Leuven: Peeters Publishers. Corpus Scriptorum Christianorum Orientalium. Scriptores Iberici 25.

Reul, C.; Christ, D.; Hartelt, A.; Balbach, N.; Wehner, M.; Springmann, U.; Wick, C.; Grundig, C.; Büttner, A.; Puppe, F. (2019). "OCR4all–An Open-Source Tool Providing a (Semi-) Automatic OCR Workflow for Historical Print-

ings". *Applied Sciences*, 9(22), 4853, 1-30. https://doi.org/10.3390/app9224853.

Schmidt, A.B.; Kindt B. (2021). "Eine syrische Amulettrolle mit Beschwörungen für Frauen: Erevan, Matenadaran, rot. syr. 72. Teil II. Wortindex". Ishac, E.A.; Csanády, Th.; Zammit Lupi, Th. (eds), *Tracing Written Heritage in a Digital Age*. Wiesbaden, 59-76.

Stone, M.E. (2021). *The Genesis Commentary by Step'anos of Siwnik' (dub.)*. Leuven: Peeters Publishers. Corpus Scriptorum Christianorum Orientalium. Scriptores Armeniaci 32.

Stone, M.E.; Kouymjian, D.; Lehmann, H. (2002). *Album of Armenian Paleography*. Aarhus: Aarhus University Press.

Vidal-Gorène, C.; Dupin, B.; Decours-Perez, A.; Riccioli, T. (2021). "A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-Resourced Languages". Lladós, J.; Lopresti, D.; Uchida, S. (eds), "Document Analysis and Recognition – ICDAR 2021. ICDAR 2021". *Lecture Notes in Computer Science*, 12823, 507-22. Cham: Springer. https://doi.org/10.1007/978-3-030-86334-0_33.

Vidal-Gorène, C.; Khurshudyan, V.; Donabédian-Demopoulos, A. (2020). "Recycling and Comparing Morphological Annotation Models for Armenian Diachronic-Variational Corpus Processing". *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 90-101. Barcelona: International Committee on Computational Linguistics (ICCL). https://aclanthology.org/2020.vardial-1.9.

Vidal-Gorène, C.; Kindt, B. (2020). "Lemmatization and POS-Tagging Process by Using Joint Learning Approach. Experimental Results on Classical Armenian, Old Georgian, and Syriac". *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*. Marseille: European Language Resources Association (ELRA), 22-7. https://aclanthology.org/2020.lt4hala-1.4.

Zohrab (Zōhrapean), Y. (1805). *Astuacašunč' matean: hin ew nor ktakaranac'*. i Venētik: I gorcarani srboyn Łazaru.