Describing Language Variation in the Colophons of Armenian Manuscripts

Emmanuel Van Elverdinghe, Bastien Kindt
Centre d’études orientales – Institut orientaliste de Louvain (CIOL), Institut des civilisations, arts et lettres (INCAL), UCLouvain, Louvain-la-Neuve, Belgium
{emmanuel.vanelverdinghe,bastien.kindt}@uclouvain.be

Abstract

The colophons of Armenian manuscripts constitute a large textual corpus spanning a millennium of written culture. These texts are highly diverse and rich in terms of linguistic variation. This poses a challenge to NLP tools, especially considering the fact that linguistic resources designed or suited for Armenian are still scarce. In this paper, we deal with a sub-corpus of colophons written to commemorate the rescue of a manuscript and dating from 1286 to ca. 1450, a thematic group distinguished by a particularly high concentration of words exhibiting linguistic variation. The text is processed (lemmatization, POS-tagging, and inflectional tagging) using the tools of the GREgORI Project and evaluated. Through a selection of examples, we show how variation is dealt with at each linguistic level (phonology, orthography, flexion, vocabulary, syntax). Complex variation, at the level of tokens or lemmata, is considered as well. The results of this work are used to enrich and refine the linguistic resources of the GREgORI project, which in turn benefits the processing of other texts.

Keywords: Ancient Armenian, Middle Armenian, colophons, lemmatization, POS-tagging, inflectional tagging, language variation

1. Preliminary notes and aims

1.1 The colophons of Armenian manuscripts

In the traditional sense, a colophon is a record of completion of a book by its scribe. The Armenian concept of yišatakar (literally “memorial”, usually translated as colophon), has a broader meaning, encompassing practically all significant annotations in manuscripts besides scholia or glosses, including personal notes left by later owners or readers. Colophons are an important part of the Armenian literary culture, where they are recognized as a full-fledged genre. As a result, they have attracted the interest of scholars for a long time, besides scholia or glosses, including personal notes left by later owners or readers. Colophons are an important part of the Armenian literary culture, where they are recognized as a full-fledged genre. As a result, they have attracted the interest of scholars for a long time, but especially since 1950, when the first systematic collection of colophons appeared in print. Since then, most colophons written until 1500 have been published in these dedicated collections, as well as colophons from the period 1601–1660.

This paper deals with a particular sub-corpus of non-scribal colophons recording the rescue of a manuscript, usually from the hands of Muslim captors. Using the abovementioned printed collections (Xačikyan, 1950, 1955, 1967; Mat’evosyan, 1984; Xačikyan, Mat’evosyan, and Lazarosyan, 2018, 2020), we identified 46 such colophons in the period leading up to 1450. The earliest of them was written in 1286; however, in several cases, the exact date is unknown and an approximate dating has been inferred. The text of these colophons was extracted from the corpus of Armenian colophons maintained at the UCLouvain and lemmatized according to the principles of the GREgORI Project (Coulie, Kindt, Kepeklian, and Van Elverdinghe, 2022). The main corpus of Armenian colophons currently comprises 1,232,652 tokens (Table 2, section A).

1.2 Language variation in Armenian

Variation affects all areas of language, occurring at the phonetical, morphological, lexical, syntactic, semantic, and pragmatic levels, and is mainly expressed across four dimensions: diachronic, diatopic, diastatic, and diaphasic (Auer and Schmidt, 2010: 226–228). The present contribution focuses on phonetical, morphological, and lexical variation in Armenian colophons within the diachronie, diatopic, and diaphasic dimensions. Proper names (anthroponyms and toponyms) are not considered here: the problems posed by this very abundant and versatile category ought to be considered separately. Upon manual inspection, the sub-corpus was found to contain an estimated 473 anthroponyms, 7 patronyms, and 82 toponyms, adding up to a provisional total of 562 tokens, or 9.62% of all tokens in the sub-corpus (see Table 1).

This percentage is almost doubled if one considers unique tokens instead of all tokens (18.30%).

Table 1: Quantitative assessment (estimation) of proper nouns in the sub-corpus

|                        | Tokens | Unique tokens |
|------------------------|--------|---------------|
| Anthroponyms (N+Ant)   | 473    | 345           |
| Patronyms (N+Pat)      | 7      | 7             |
| Toponyms (N+Top)       | 82     | 71            |
| Proper nouns total     | 562    | 421           |
| As percentage of sub-corpus | 9.62% | 18.30%        |

The high variability and unpredictability of these categories creates a serious challenge. As an example, the only attestation of the name Երեւան Erewan in the sub-corpus, does not refer to the current capital of Armenia, but to an elderly priest. But the main difficulty with processing a proper noun lies in formulating an adequate lemma, owing to the number of different variants, spellings, and paradigms attested in the texts. For instance, the name George appears variously in the sub-corpus as Գորգ, Գեորգ, Գեորգեոս, and Գեորգ. In addition to such true variants, there is the widespread issue of scribal inconsistency, which cannot always be easily resolved. In the following case,
one colophon has as many as four different spellings for the same toponym: Ծինայանի` Ծինանուն, Ծինավանի` Ծինավան, Ծինենավանի` Ծինենավան, and Ծինիվանի` Ծինիվան. These questions, however interesting, outstretch the aims of this paper and should be dealt with at a later stage. Specific studies have been devoted to various aspects of linguistic variation in Armenian colophons, focusing principally on the period from the twelfth to the fifteenth century: sound change (Harut’yunyan, 2014b), diachronic morphology (Harut’yunyan, 2014a; Hovsep’yán, 1997), dialectal features (Jahykyan, 1997), neologisms (Margaryan, 1993), anthroponymy (Harut’yunyan, 2018a, 2018b; Weitenberg, 2005), and stylistic patterns (Van Elverdinghe, 2018, 2022). Obviously, these developments of the Middle Armenian idiom are not specific to colophons. Most of them have been described by Karst (1901), drawing from literary, legal, medical, etc., texts. Since then, numerous studies have enriched our knowledge of Middle Armenian sources. The sub-corpus studied here was selected because it shows a more diverse linguistic picture than a random sampling of Armenian colophons of the same period would. This is due to the fact that many colophons of this group are not written by professional scribes and do not follow the customs and patterns of colophon writing. Therefore, the widespread tendency to normalize and conformity to the rules of Classical Armenian recedes, while the spoken Middle Armenian idiom infiltrates the written medium. This allows for more or less considerable linguistic variation within each colophon.

1.3 Linguistic resources of the GREgORI Project

The automated analysis of this sub-corpus of Armenian colophons was carried out using tools and linguistic data of the GREgORI Project. The Armenian language shares characteristics of both inflected and agglutinative languages. As such, inflected simple forms can receive prepositional suffixes as well as determinative suffixes. In their current state, the linguistic resources of the GREgORI Project consist of a set of 315,952 simple word-forms (i.e. inflected words such as աշխատողաց աշխատողաց), on the one hand, and a set of 883,171 polylexical forms (such as աշխատողաց, i.e. աշխատողաց-ն), on the other hand. Together, these two sets totalize 1,199,123 tokens, simple or polylexical, which are recorded along with 30,311 lemmata (lexical entries) and the corresponding part-of-speech of these lemmata. Word-forms are either taken from the corpora already processed in the past or generated automatically (under human supervision) in order to improve, as much as possible, the lexical coverage during the processing of new corpora. The sum of these data constitutes a reference lexicon (Coulie, Kindt, Kepeklian, and Van Elverdinghe, 2022). On that basis, the main goals of the GREgORI Project can be reached, viz to provide scholars with tagged corpora, lemmatized concordances or indexes, and online, searchable corpora.

2. Processing and preliminary evaluation

The processing phase consists in lemmatization, POS-tagging, and inflectional tagging. It is subdivided in three steps, as described by Kindt, Vidal-Gorène, and Delle Donne (2022; Vidal-Gorène and Kindt, 2022): 1) analysis by lexical look-up, matching the vocabulary of the corpus with the data gathered in the reference lexicon; 2) analysis using an RNN model; 3) manual check of the analysed data. Only then can scholars be provided with a final, tagged corpus. The first step ensures a highly accurate tagging, but fails to identify unknown words and does not solve lexical ambiguities. The second step resort to an RNN model previously trained with already processed corpora of the GREgORI Project and applied by Calña to the study of new corpora. In that case, the outcomes are complete, since the process does not disregard unknown words and resolves lexical ambiguity. However, they remain statistical predictions, and not analyses grounded on a common linguistic approach. A considerable advantage to this hybrid approach is that it alleviates the human intervention necessary during the third step, before the final data can be delivered (Kindt, Vidal-Gorène, and Delle Donne, 2022; Vidal-Gorène and Kindt, 2020).

A PDF version of the lemmatized concordance of the sub-corpus is available on the GREgORI website1. The sub-corpus is also available on the online interfaces of the GREgORI Project2.

![Table 2: Number of tokens and unique tokens in the Armenian colophons (main corpus and sub-corpus)](https://www.gregoriproject.com)

Table 2: Number of tokens and unique tokens in the Armenian colophons (main corpus and sub-corpus)

| Section A | Main corpus of Armenian colophons |
|-----------|----------------------------------|
| Tokens | 1,232,652 |
| Unique tokens | 144,347 |

| Section B | Sub-corpus of Armenian colophons |
|-----------|----------------------------------|
| Tokens | 5,845 |
| Unique tokens | 2,300 |

| Step 1 – Analysis by lexical look-up |
|-----------------------------------|
| Lemma = 0 | 1,263 |
| Lemma = 1 | 3,281 |
| Lemmata > 1 | 1,301 |

| Step 2 – Analysis using an RNN model |
|-----------------------------------|
| Lemma = 1 | 5,845 |

| Step 3 – Checking results (April 2022) |
|-----------------------------------|
| Already checked | 4,381 |

---

1. https://uclouvain.be/fr/instituts-recherche/incal/ciol/ gregori-project.html

2. https://www.gregoriproject.com
tokens with no lemma, with one lemma, with more than one lemma). For the reasons explained above, the results obtained by RNN (step 2) are equal to the total number of words. Finally (step 3), the current number of already checked results is given.

The lexical analysis of Armenian colophons (main corpus or sub-corpus) is still a work in progress. Most notably, the analysis and lemmatization of proper nouns has been deferred to a later date (see above, 1.2). Nonetheless, the current results already allow using lemmata, POS-tags and inflectional analysis to explore adequately the sub-corpus under consideration. Indeed, tagged data are very helpful in order to describe language variation in the sub-corpus and to single out relevant examples. Many of the 1.263 unknown words (Lemma = 0) highlighted during step 1 (see table 2) are examples of linguistic variation: they bear witness to non-classical strata of the Armenian language that are not yet fully described in the linguistic resources of the GREgORI Project.

3. Selected examples of linguistic variation

The following examples are organized according to the linguistic level at which they occur. They are meant as a representative sample of the different phenomena attested in the corpus, and of their description in the linguistic resources of the GREgORI Project. The issue of which dialect, period, etc., is affected by these variations is too complex to be dealt with here. The same goes for the precise linguistic constraints surrounding these changes. All these examples concern words for which the resources of the GREgORI Project fail to offer an analysis, counted in the 1.263 unknown words (“lemma = 0”) quoted in table 1 (step 1).

3.1 Phonology

At the phoneme level, the language of colophons reflects the general evolution of the Armenian vocalic system, including monophthongization and merger of some sounds (except at the beginning of words), such as: aw (also written หาร or หาร (1), หาร or หาร (2)). Consonants are subject to multiple variations, among which one can cite, in addition to the well-known consonant shift affecting a number of dialects, the devoicing and aspiration of voiced consonants in certain contexts (3), and the devoicing of final deictic sounds (except at the beginning of words), such as:

1) H14 681, p. 546 l. 6: סוּשָׂפָ ‘ol-ac’ (teuch-AGN-GEN/DAT/ABL.PL) “handlers” (Cl. aw = unami aw וםשפא ac’)

2) H15A 699, p. 619 l. 37: գնութ կ ’որեն “wheat” (Cl. գնութ, գութ, N+Com.՝AwkGp)

3) H14 685, p. 549 l. 20: աւագ awak ‘greater, senior” (Cl. աւաց awag)

4) H15A 616, p. 543 l. 6: ւայագ awawt ’k’-t (prayer-NOM.PL-that) “your prayers” (Cl. աւաց awawt k’d)

3.2 Orthography

These sound changes in turn gave rise to incorrect or hypercorrect spellings. For instance, the medieval letter .Should, which stands for the old diphthong aw in positions where the latter was monophthongized, is also incorrectly used where aw was actually realized as /aw/ (5). Another orthographical feature is that the ephenthic schwa is occasionally written in positions where, according to the spelling rules of Classical Armenian, it should not appear (6).

5) H14 685, p. 549 l. 19: օւնտունչ դետարան-s (gospel-this) “this Gospel [book]” (Cl. aw = unami օւնտունչ awetaran)

6) H14B 799, p. 447 l. 10: վերստին verastin “once again” (Cl. վերստի aw-verstiv)

3.3 Declension

A number of words undergo paradigmatic reorganization, changing from one thematic paradigm to another (7) or, in the case of irregular paradigms, switching to a regular, thematic paradigm (8; 9). In parallel, several new endings develop, notably plurals in -սեր (10) and locatives in -սեր (11).

7) H15A 129, p. 128 l. 38: պասուիկ-ա (service-INSTR.PL) “with [liturgical] vessels” (Cl. պասուիպ aspuk)

8) H13 478b, p. 595 l. 12: փունրա p’ok ’r-i (small-GEN/DAT/LOC.SG) “small” (Cl. փունրա p’ok’u)

9) H15A 38, p. 40 l. 34: զվաների z-van-en-r (DOBJ-monastery-PL-the) “the monasteries” (Cl. զվաներ zvansn)

4 For further information about these linguistic phenomena, the reader is referred to the works cited above (1.2).

10) H15C 544, p. 403 l. 28: քուերան k uer-ac ’ (sister-GEN/DAT/ABL.PL) “sisters” (Cl. քուերա n’kær’c)

11) H14 676, p. 543 l. 19: եւիական-եւիական i handerj-el-um-n (in prepare-PART-LOC-the) “in the future”

3.4 Conjugation

Similar evolutions characterize the verbal system. Monosyllabic third person singular aorist forms receive an augment in -եր- or -եր- if they already had an
augment in Classical Armenian. The latter evolution applies, among others, to verb *tam* “to give”, which even gets a whole new aorist paradigm (13). An important element in the reconfiguration of the verbal system is the emergence of a particle *ku* (ko / k- ) to mark the indicative mood (14).

H14 679a, p. 544 l. 35: քուրփ է-զարک (AOR.3.SG-strike) “[the khan] struck” (Cl. քուրփ զարկ)

13) H15A 418b, p. 392 l. 21: սուր-ի տուի (give-AOR.1.SG) “I gave”
   = սուրեն.մուտ.V:ԷԻ3s (Cl. լանու. ետու)

14) H14B 670, p. 295 l. 32: պարոն քու-ար (IND-want-IMPFT.3.SG) “[the sultan] wanted”
   = պարոն և.1+Prep@/noclq@,n.qla:V:ԷԻ3s

3.5 Vocabulary

The vocabulary of colophons includes words not found in classical texts, such as dialectal or colloquial words (15), neologisms (16), and loan-words (17; 18). Purely semantic variation are, as a rule, not recorded by the GREORI project.

15) H14B 670, p. 295 l. 6: ախատ աչք (plunder-AOR.3.SG) “he plundered”
   = աչքը,պարոն:V:ԷԻ3s

16) H15A 1* p. 3 n. 1 1: երետ նելաչ ուի “slant-eyed”, from երետ ու (narrow) and աչք աչք “eyes”
   = երետու.Ա

17) H14 593d, p. 484 l. 32: հապատ հալալ “legitimate”, from Arabic Հալալ halal
   = հապատ Ա

18) H14 681, p. 546 l. 16: պարոն պարոն “sir”, from French baron
   = պարոն N+Com

3.6 Syntax

The syntax of colophons shows a number of peculiarities, some of which are common to other Middle Armenian literary texts. As an example, one can cite the fact that the nominative plural ending -k- is increasingly used for the direct object, instead of the accusative plural ending -s (especially with *pluralia tantum*) (19).

19) H14B 670, p. 296 l. 1: կատարեի եզեր զավջում նա կակ-ո “he fulfilled the sultan’s wish” (fulfil-AOR.3.SG DOBJ-sultan-GEN-SG-the will-NOM.PL-the)
   = կատարեւք,կատարեւք:V:ԷԻ3s q.q1+Prep@/nqna:V,unqna:V,unqna:V+N+Com@/l,li:PRO+Dem

4. Complex variation

In some of the examples given above, more than one feature can be ascribed to linguistic variation. Thus in (14), not only is the particle քо- an innovation, but the verbal lemma itself, եզեր “to want”, is a Middle Armenian variant of the classical verbal կատարել yuzem “to seek”, in which a sound change (loss of the initial glide) coincides with semantic evolution. Likewise, some lemmata concentrate different instances of variation, as lemmatized concordances readily show.

Appendix 9.1 lists the attested tokens of the lemma բերդ, one of three words with the meaning of “fortress, castle” in the sub-corpus (the other two being ամրոց and կլա). The words բերդ, բերդուն, բերդուներ, and զբերդուն demonstrate the plural formation in -(n)er (9)—notice how not a single classical plural form of this lemma is found in the sub-corpus—, while պլու(ր)եր հերռի is a case of devoicing and aspiration of a voiced consonant after r (3).

Appendix 9.2 presents a concordance of the lemma տված “to give”, showing several non-classical forms of the active aorist paradigm (13): first person singular տված ուի, third person singular տված երետ and տված էրետ (12), and first person plural տված տված “twink” (6) and տված տված “twav”. In addition, the sub-corpus contains an occurrence of the Middle Armenian participial form տվայլ tvac, appearing as part of a periphrastic past tense.

5. Conclusion

The corpus of Armenian colophons constitutes an invaluable collection of texts, both historically and linguistically (Harut’yunyan, 2019; Stone, 1995; etc.). The language of this corpus stands out for its diachronic, diatopic, and diaphasic variation. Therefore, a systematic analysis of the vocabulary of colophons using NLP tools will be helpful to increase our knowledge and understanding of the varieties, evolution, and uses of the Armenian language.

Already before the sub-corpus discussed here was processed, the resources of the GREORI Project had been used on the whole corpus to facilitate an investigation into the formulaic patterns that characterize the style of Armenian colophons (Van Elverdinghe, 2018, 2022). Lemmatization, POS-tagging, and inflectional tagging of the corpus make it possible to successfully execute complex search queries, such as is required to detect and analyse speech patterns.

The long-term goal is to achieve full lemmatization of the whole corpus of Armenian colophons; in the meantime, applications on more limited sub-corpora like the one under consideration here are expected. Enriching the linguistic resources of the GREORI Project with forms found in colophons also represents a step forward towards the treatment of other Middle Armenian texts, especially texts of a documentary nature, such as inscriptions, of which there is already an example on the GREORI website (Goepf, Mutafian, and Ouzounian, 2012).

As regards the processing of proper nouns, two avenues could be explored. One relies on manual lemmatization of newly encountered forms, basing the decisions on reference works such as the dictionaries by Açağyan (1942–1962) for anthroponyms and by Hakopyan, Melik-BAŞXYAN, and Barselyan (1986–2001) for toponyms. The other path entails complete or partial automation of the initial process using an existing data-set. Unfortunately, any corpus designed for modern Eastern Armenian, such as pioNER (2018 – see Ghukasyan et al., 2018), can hardly be exploited from a Classical or Middle Armenian perspective. The most appealing prospect at this point is the ongoing digitization and full OCR of Adjarian’s Dictionary of Armenian Personal
A number of annotated corpora are already freely available on the web, such as Arak-29 (since 2002) for Classical Armenian (mainly) or EANC (2006–2009) for Modern Eastern Armenian. Nevertheless, Ancient Armenian, generally speaking, remains an under-resourced language. Corpora featuring high-quality lexical tagging and available through interoperable formats are still scarce (Vidal-Gorène and Decours-Perez, 2020; Vidal-Gorène and Kindt, 2022). By processing this corpus, the GREGori Project, in close connection with Calìf and the UCLouvain, intends to build up its linguistic resources and tailor them to the particular idiom of colophons, a task which is not only essential for a successful study of this textual content, but also paves the way for future research on other medieval Armenian sources.

6. Acknowledgements
Emmanuel Van Elverdinghe is a Postdoctoral Researcher of the Fonds de la Recherche Scientifique – FNRS. Special thanks are due to Professor Bernard Coullie (UCLouvain), to Gabriel Kepeklian (UCLouvain), who manages the GREGori Project’s Armenian resources, and to Chahan Vidal-Gorène (Calìf), who is responsible for implementing an RNN model trained with the data of the GREGori Project.

7. Bibliographical References
Açaryan, H. (1942–1962). Hayoc’ anjnanunneri bararan [= A Dictionary of Armenian Personal Names], 5 vols. Erevan: Petakan hamalsarani htrarakač’ut’yun.
Auer, P. and Schmidt, J. E. (Eds.). (2010). Language and Space: An International Handbook of Linguistic Variation. Volume 1: Theories and Methods. Berlin, New York: De Gruyter Mouton.
Coullie, B., Kindt, B., Kepeklian, G., and Van Elverdinghe, E. (2022). Etiquettes morphosyntaxiques et flexionnelles pour le traitement automatique de l’arménien ancien. Le Muséon, 135(1–2): 209–241.
Ghukasyan, Ts. et al. (2018), pioNER: Datasets and Baselines for Armenian Named Entity Recognition. In A. Avetisyan et al. (Eds.), 2018 Ivanovik Ips Ras Open Conference, Dedicated to the 70th Anniversary of Computer Science in Russia. ISPRAS 2018, 22–23 November 2018, Moscow, Russia, Federation: Proceedings, pp. 56–61. Los Alamitos, Washington, and Tokyo: IEEE Computer Society Conference Publishing Services.
Goepf, M., Mutafian, C., and Ouzounian, A. (2012). L’inscription du régent Constantin de Papeрон (1241). Redécouverte, relecture, remise en contexte historique. Revue des études arméniennes, 34: 243–287.
Hakobyan, T. ‘X., Melik’-Başxyan, St. T., and Barse-lyan, H. X. (1986–2001). Hayastani ev harakic’ šrjnan-neri telanunn-neri bararan [= Dictionary of Toponymy of Armenia and Adjacent Territories], 5 vols. Yerevan: Erevani hamalsarani htrarakač’ut’yun.
Harut’yunyan, X. (2014a). Holovman hamakarga XV dari hayeren jerageri hıšatarkeranerum. Banber Erevani Hamalsarani. Hasarakakan Gitut’yunner. Săr’k’ 2. Banasirat’yun [= The System of Declension in the Colophons of Armenian Manuscripts of the XV Century. Bulletin of Yerevan University. Social Sciences. Volume 2: Philology], 5(143): 123–131.
Harut’yunyan, X. (2014b). XIV–XV dareri hayeren jerageri hıšatarkeranerem lezvi hıç’yunakan himna-kan biut’agiro. Banber Matenadarani [= An Elementary Phonetic Description of the Language of Armenian Colophons of the 14th–15th Centuries. Bulletin of Matenadaran], 20, 175–187.
Harut’yunyan, X. (2018a). Anjnanunnera hayeren jerageri hıšatarkeranerum. 1. Norahayt anjnanunner ZA–ZG, dareri hıšatarkeranermc’. Banber Matenadaran [= Personal Names in the Colophons of Armenian Manuscripts. 1: Newfound Personal Names in Colophons of the 11th–13th Centuries. Bulletin of Matenadaran], 25: 187–217.
Harut’yunyan, X. (2018b). Del armatov baladrav anjnanunnera vinagrumer ev jerageri hıšatarkeranerum. In Eritasardakan 3-rd gitulzolov zeuk’cunner (Erevan, 2017 t., noyember 28-30) [= Personal Names with the Root del in Armenian Colophons and Inscriptions. In Papers of the III Youth Conference (Erevan, 2017, November 28-30)], pp. 115–133. Yerevan: Armav htrarakač’ut’yun.
Harut’yunyan, X. A. (2019). Hayaren jerageri hıšatarkeranera [= The Colophons of Armenian Manuscripts]. Yerevan: Matenadaran.
Hovsep’yán, L. S. (1997). ŽG dari hayeren jerageri hıšatarkeranerum lezuni [= The Language of the Armenian Colophons of the 13th Century]. Yerevan: «Van Aryan» htrarakač’ut’atun.
Jahukyan, G. B. (1997). Barbarayin erevuyt’nera hay-akakan hıšatarkeranerum [= Dialect Features in Ar-menian Colophons]. Yerevan: «Van Aryan» htrarakač’ut’atun.
Karst, J. (1901). Historische Grammatik des Kilikisch-Armenischen. Strassbourg: Verlag von Karl J. Trübner.
Kindt, B., Vidal-Gorène, Ch., and Delle Donne, S. (2022). Analyse automatique du grec ancien par ré-seau de neurones. Évaluation sur le corpus De Thes-salonica Capit. BABELAO, 10–11: 537–562.
Margaryan, Al. S. (1993). Norahayt bıut’agiro hayeren jerageri XIV–XV dd. hıšatarkeranerum. Patma- banasirakan handes [= Newfound Words in the Colophons of Armenian Manuscripts of the 14th– 15th Centuries. Historical-Philological Journal], 1–2(137–138): 35–42.
Mat’evoys, A. S. (1984). Hayeren jerageri hıšatarker-aner. ŽGdar [= Colophons of Armenian Manuscripts: 13th Century]. Yerevan: Haykakan SSH Gitut’yunneri Akademiyai htrarakač’ut’yun.
Stone, M. E. (1995). Colophons in Armenian Manuscripts. In E. Condello and G. De Gregorio (Eds.), Scribi e colofoni. Le sottoscrizioni di copisti dalle origini all’avvento della stampa. Atti del seminario di Erice. X Colloquio del Comité international de paléographie latine (23–28 ottobre 1993), pp. 463–471. Spoleto: Centro italiano di studi sull’alto me-dioevo.
Van Elverdinghe, E. (2018). Recurrent Pattern Modelling in a Corpus of Armenian Manuscript Colophons. Journal of Data Mining and Digital Human-
ties, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages: 8 pp.
Van Elverdinghe, E. (2022). Modèles et copies. Étude d’une formule des colophons de manuscrits arméniens, VII–XIX siècles. Louvain: Peeters.
Vidal-Gorène, Ch. and Decours-Perez, A. (2020). Languages Resources for Poorly Endowed Languages: The Case Study of Classical Armenian. In N. Calzolari et al. (Eds.), LREC 2020, Marseille. Twelfth International Conference on Language Resources and Evaluation, May 11-16, 2020, Palais du Pharo, Marseille, France: Conference proceedings, pp. 3145–3152. Paris: The European Language Resources Association (ELRA).
Vidal-Gorène, Ch. and Kindt, B. (2020). Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old Georgian, and Syriac. In R. Sprugnoli and M. Passarotti (Eds.), 1st Workshop on Language Technologies for Historical and Ancient Languages, (LT4HALA 2020): Proceedings, pp. 22–27. Paris: European Language Resources Association (ELRA).
Vidal-Gorène, Ch. and Kindt, B. (2022). From manuscript to tagged corpora. An automated process for Ancient Armenian or other under-resourced languages of the Christian East. Armeniaca, 1 (submitted).
Weitenberg, J. J. S. (1995). The Role of Morphologic Variation in Medieval Armenian Poetry. In J. J. S. Weitenberg (Ed.), New Approaches to Medieval Armenian Language and Literature, pp. 121–134. Amsterdam, Atlanta (GA): Rodopi.
Weitenberg, J. J. S. (2005). Cultural Interaction in the Middle East as Reflected in the Anthroponomy of Armenian 12th – 14th Century Colophons. In J. J. van Ginkel, H. L. Murre-van den Berg, and Th. M. van Lint (Eds.), Redefining Christian Identity: Cultural Interaction in the Middle East since the Rise of Islam, pp. 265–263. Louvain, Paris, Dudley (MA): Uitgeverij Peeters; Departement Oosterse Studies.
Xač’ikyan, L., Mat’evosyan, A., and Łazarosyan, A. (2018). Hayeren jeṙagreri hišatakaranner. ŽD dar. Masn A (1301-1325 t’t’.) [= Colophons of Armenian Manuscripts: 14th Century. Part 1 (1301–1325)]. Yerevan: «Nairi» hratarakč’ut’yun.
Xač’ikyan, L., Mat’evosyan, A., and Łazarosyan, A. (2020). Hayeren jeṙagreri hišatakaranner. ŽD dar. Masn B (1326-1350 t’t’.) [= Colophons of Armenian Manuscripts: 14th Century. Part 2 (1326–1350)]. Yerevan: Matenadaran.
Xač’ikyan, L. S. (1950). ŽD dari hayeren jeṙagreri hišatakaranner [= Colophons of Armenian Manuscripts of the 14th Century]. Yerevan: Haykakan SSH Gitut’yunneri Akademiayi hratarakč’ut’yun.
Xač’ikyan, L. S. (1955). ŽE dari hayeren jeṙagreri hišatakaranner. Masn aṙaǰin (1401—1450 t’t’.) [= Colophons of Armenian Manuscripts of the 15th Century. Part One (1401–1450)]. Yerevan: Haykakan SSH Gitut’yunneri Akademiayi Hratarakč’ut’yun.
Xač’ikyan, L. S. (1967). ŽE dari hayeren jeṙagreri hišatakaranner. Masn errord (1481—1500 t’t’.) [= Colophons of Armenian Manuscripts of the 15th Century. Part Three (1481–1500)]. Yerevan: Haykakan SSH Gitut’yunneri Akademiayi Hratarakč’ut’yun.

8. Language Resource References
Arak-29, (since 2002). Առակ-29 / Arak-29. Առակ-29կրթամշակութային հիմնադրամ, https://arak29.org.
Calza, (since 2014). Calza, https://calza.fr.
EANC. (2006–2009). Eastern Armenian National Corpus. Corpus Technologies, http://www.eanc.net.
GREgORI Project. (since 1990). GREgORI – Software, linguistic data and tagged corpora for ancient GREek and ORiental languages, https://uclouvain.be/fr/instituts-recherche/incaiol/gregori-project.html, ISSN 2736-7657 (Bernard Coulie, Academic supervisor).
pioNER. (2018). pioNER - named entity annotated datasets and GloVe models for the Armenian language, https://github.com/ispras-texterra/pioneer.
9. Appendix: samples of concordances

9.1  Concordance of the lemma թամ բերդ (fortress) in the sub-corpus

| lemma | XIV_B 670 0 296 9 | XIV_B 670 0 296 8 | XIV_B 670 0 296 2 | XIV_A 347 0 328 11 | XIV_B 670 0 295 28 | XIV_A 580 0 515 10 | XIV_A 330 0 314 8 | XIV_B 670 0 295 24 |
|-------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| lemma | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ |
| lemma | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ |
| lemma | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ |
| lemma | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ | թամ բերդ |

9.2  Concordance of the lemma մաու տամ (to give) in the sub-corpus

| lemma | XIV_A 437 3 492 26 | XIV_A 111 3 121 8 | XIV_A 699 0 619 37 | XIV_A 681 0 546 11 | XIV_A 1 0n1 3n 13 | XIV_A 585 2 519 5 | XIV_A 699 0 619 38 | XIV_A 585 1 518 21 |
|-------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| lemma | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ |
| lemma | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ |
| lemma | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ |
| lemma | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ | մաու տամ |