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 *yisatakar* (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, 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 (Xač’ikyan, 1950, 1955, 1967; Mat’evosyan, 1984; Xač’ikyan, 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 diachronic, 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>
<thead>
<tr>
<th>Tokens</th>
<th>Unique tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthroponyms (N+Ant)</td>
<td>473</td>
</tr>
<tr>
<td>Patronymics (N+Pat)</td>
<td>7</td>
</tr>
<tr>
<td>Toponyms (N+Top)</td>
<td>82</td>
</tr>
<tr>
<td>Proper nouns total</td>
<td>562</td>
</tr>
<tr>
<td>As percentage of sub-corpus</td>
<td>9.62%</td>
</tr>
</tbody>
</table>

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

The high variability and unpredictability of these categories creates a serious challenge. As an example, the only attestation of the name Երեւան Erevan 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 Գեորգ, Գէորգ Gēorg, Գէորգէոս Gēorgēos, and Գորգ Gorg. 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: Բուժառապարգություն Տունավանից’, Բուժառապարգություն Տունեվանից’, Բուժառապարգություն Տունիվանից’ and Բուժառապարգություն Տունիվանից’. 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; Hosep’yan, 1997), dialectal features (Jahukyan, 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.

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. This is due to the fact that many colophons of this group are not written by professional scribes and 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 normalization 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 աշխատողացն զաշխատողաց ), on the one hand, and a set of 883,171 polylexical forms (such as զաշխատողաց ’-n), 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 and 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 resorts to an RNN model previously trained with already processed corpora of the GREgORI Project and applied by Calfa 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 website. The sub-corpus is also available on the online interfaces of the GREgORI Project.

<table>
<thead>
<tr>
<th>Section A</th>
<th>Main corpus of Armenian colophons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>1,232,652</td>
</tr>
<tr>
<td>Unique tokens</td>
<td>144,347</td>
</tr>
</tbody>
</table>

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

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 ø) > ø (1), ø > e, ev(y) > õ (= e) (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 -d (4).

1) H14 681, p. 546 l. 6: հոտալ -ot-ac’ (toucb-AGN-GEN/DAT/ABL.PL) “handlers” (Cl. շակաբ-ազնշաղ շակված-դակ’)
   = լուսազնված,լուսազնված,N+Com:ArPnPp
2) H15A 699, p. 619 l. 37: ընտեղ կ’օրեն “wheat” (Cl. գնուցու կ’օրեն) = գնուցու,գնուցու,N+Com:AsNsUs
3) H14 685, p. 549 l. 20: ավաք ավաք “greater, senior” (Cl. ավաք-ավաք awag)

#### 3.2 Orthography

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

4) H15A 616, p. 543 l. 6: աղաւթք aławt’k’-t (prayer-NOM.PL-that) “your prayers” (Cl. աղաւթք aławt’k’)
   = աղաւթք,աղաւթք,N+Com:Np@u,u,PRO+Dem

#### 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 -ner (10) and locatives in -um (11).

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

5) H14 685, p. 549 l. 19: աղաւթք aławt’k’ (gospel-ths) “this Gospel” (Cl. աղաւթք aławt’k’)
   = աղաւթք,աղաւթք,N+Com:AsNs@u,u,PRO+Dem
6) H14B 799, p. 447 l. 10: աղաւթք aławt’k’ “once again” (Cl. աղաւթք aławt’k’)
   = աղաւթք,աղաւթք,I+Adv

#### 3.4 Conjugation

Similar evolutions characterize the verbal system. Monosyllabic third person singular aorist forms receive an augment in e- or é- (12), or er- 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).

(12) H14 679a, p. 544 l. 35: կտավը է-զարկ (AOR.3.SG-strike) “[the khan] struck” (Cl. կար զարկ)
= կտավը, կարզարկավ: V:EI13

(13) H15A 418h, p. 392 l. 21: ունի տի-ի (give-AOR.1.SG) “I gave”
= ունի, տի-ի: V:EI13s կամ (l. unu, etu)

(14) H14B 670, p. 295 l. 32: ռազմական k- uz- էր (IND-want-IMPF.3.SG) “[the sultan] wanted”
= ռազմական, k- uz- էր: V:EI13s

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 variations are, as a rule, not recorded by the GRE ORI project.

(15) H14B 670, p. 295 l. 6: ջիհամբ նիշ-էկ (plunder-AOR.3.SG) “he plundered”
= ջիհամբ, նիշ-էկ: V:EI13s

(16) H15A 1*, p. 3 n. 1 l. 5: նեղաչուն na- է է (slant-eyed), from նեղ ոտ “narrow” and է է (Cl. ոտ է է) “eyes”
= նեղաչուն, ոտ է է: A

(17) H14 681, p. 546 l. 16: տամ պարոն “sir”, from French baron
= տամ, 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: քրանունութք քրանունութք քրանունութք կատա-էկ’ z-sultan-i-n kam-k’n “he fulfilled the sultan’s wish” (fulfil-AOR.3.SG DOBJ-sultan-GEN.SG-the will-NOM.PL-the)
= քրանունութք, քրանունութք, քրանունութք, կատա-էկ’ z-sultan-i-n kam-k’n: V:EI13s
q.4.+Prep@ ւմի.զունուի, և, ն.զունուի.Ն+Com @l, ւ, PRO+Dem քրանունութք (քրանունութք): N+Com @l, ւ, 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 կ- an innovation, but the verbal lemma itself, ն.զունու աներ “to want”, is a Middle Armenian variant of the classical verb եւրահ zyzen “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 տկա “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 կու (ko / k-) to mark the indicative mood (14).

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 variations are, as a rule, not recorded by the GRE ORI project.

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).

The corpus of Armenian colophons constitutes an invaluable collection of texts, both historically and linguistically (Harutyunyan, 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 GRE ORI 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 GRE ORI 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, as such as inscriptions, of which there is already an example on the GRE ORI 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 Acharian (1942–1962) for anthroponyms and by Hakobyan, Melik-Başyan, andBarselyan (1986–2001) for toponyms. The other path entails complete or partial automation of the initial process using an existing dataset. Unfortunately, any corpus designed for modern Eastern Armenian, such as pioNER (2018 – see Ghuksayan 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
Names (Açaryan, 1942–1962) by Calif, which should result in a suitable, if incomplete, dataset of anthro-
ponyms. A number of annotated corpora are already freely avail-
able on the web, such as Arak-29 (since 2002) for Classical Armenian (mainly) or EANC (2006–2009) for Modern Eastern Armenian. Nevertheless, Ancient Ar-
menian, generally speaking, remains an under-resourced language. Corpora featuring high-quality lexical tag-
g and available through interoperable formats are still scarce (Vidal-Gore and Decours-Perez, 2020; Vidal-Gore and Kindt, 2022). By processing this corpus, the GREgORI Project, in close connection with Calif and the UCLouvain, intends to build up its lin-
guistic resources and tailor them to the particular idiom of colophons, a task which is not only essential for a suc-
cessful 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 Coulie (UCLouvain), to Gabriel Kepeklian (UCLouvain), who manages the GREgORI Project’s Armenian resources, and to Chahan Vidal-Gore (Calf), who is responsible for implementing an RNN model trained with the data of the GREgORI Project.

7. Bibliographical References
Coulie, B., Kindt, B., Kepeklian, G., and Van Elver-
dinghe, E. (2022). Etiquettes morphosyntaxiques et flexionnelles pour le traitement automatique de l’ar-
Hakobyan, T. X., Melik’-Başxyan, St. T., and Barse-
lyan, H. X. (1986–2001). Hayastani ev harakic’ šrjan-
eri telunanneri bararan [= Dictionary of Toponymy of Armenia and Adjacent Territories], 5 vols. Yere-
van: Erevani hamalsarani hratarak’e’t’yun.
Harut’yunyan, X. (2014b). XIV–XV dareri hayeren jérargeri hıšaťakaranerum lezvi hıc’yunakan hın-
man bnut’agiro. Banber Matenadarani [= An Ele-
Harut’yunyan, X. A. (2019). Hayeren jérargeri hıša-
ťakaraner = The Colophons of Armenian Manus-
cripts. Yerevan: Matenadaran.
Hovsep’yun, L. S. (1997). ŽG dari hayeren jérargeri hıšaťakaranerum lezun [= The Language of the Ar-
Jahukyan, G. B. (1997). Barbarayin erevuyt’nera hay-
kakan hıšaťakaranerum [= Dialect Features in Ar-
ménian Colophons]. Yerevan: «Van Aryan» hra-
ratarkan’atun.
Karst, J. (1901). Historische Grammatik des Kilikisch-
Kindt, B., Vidal-Gore, C., and Delle Donna, S. (2022). Analyse automatique du grec ancien par ré-
seau de neurones. Évaluation sur le corpus De Thes-
salomica Capita. BABELAO, 10–11: 537–562.
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’evosyan, A. S. (1984). Hayeren jérargeri hıša-
takaraner. ŽG dar [= Colophons of Armenian Manu-
scripts: 15th Century]. Yerevan: Haykakan SSH Git-
t’yunner Akademiayi hratarak’e’t’yun.
Stone, M. E. (1995). Colophons in Armenian Manu-
scripts. In E. Condello and G. De Gregorio (Eds.), Scribi e colofoni. Le sottoscrizioni di copisti dalle origi-
471. Spoleto: Centro italiano di studi sull’alto me-
dioevo.
Van Elverdinghe, E. (2018). Recurrent Pattern Modell-
ing in a Corpus of Armenian Manuscript Col-
phons. Journal of Data Mining and Digital Human-

ties, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages: 8 pp.

8. Language Resource References

9. Appendix: examples of concordances

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

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