*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3145–3152
Marseille, 11–16 May 2020

# Languages Resources for Poorly Endowed Languages : The Case Study of Classical Armenian

**Chahan Vidal-Gorène, Aliénor Decours-Perez**
École Nationale des Chartes-PSL, Calfa
65 rue de Richelieu (75002 Paris), MIE Bastille - 50 rue des Tournelles (75003 Paris)
chahan.vidal-gorene@chartes.psl.eu, alienor.decours@calfa.fr

**Abstract**

Classical Armenian is a poorly endowed language, that despite a great tradition of lexicographical erudition is coping with a lack of resources. Although numerous initiatives exist to preserve the Classical Armenian language, the lack of precise and complete grammatical and lexicographical resources remains. This article offers a situation analysis of the existing resources for Classical Armenian and presents the new digital resources provided on the Calfa platform. The Calfa project gathers existing resources and updates, enriches and enhances their content to offer the richest database for Classical Armenian today. Faced with the challenges specific to a poorly endowed language, the Calfa project is also developing new technologies and solutions to enable preservation, advanced research, and larger systems and developments for the Armenian language.

**Keywords:** Lexical Database, Handwritten and Typewritten Document Recognition, Less-Resourced Language, Collaborative Resource Construction & Crowdsourcing, Corpus, Digital Humanities, Multilinguality, OCR

## 1. Introduction

The Armenian language is an Indo-European language written since the V$^{\mathrm{th}}$ century, when the monk *Maštocʻ* created the alphabet in 405 AD, according to the Armenian historian *Koriwn*. Early on a rich and prolific literature arise: translation of the Bible and of Greek and Syriac patristic texts, historiographies, theological works, etc. Classical Armenian, or *grabar* (”from writing”), in a strict sense refers to the language of the V$^{\mathrm{th}}$ century literary and religious texts, and in the broad sense to the written language that develop and expand through the Middle Ages and Modern Times. Indeed, preserved as a ”reference point” (Lamberterie, 1992) the use of Classical Armenian continues on until the XIX$^{\mathrm{th}}$ century, growing richer with technical terminology taken from the Ancient Greek through the translation of secular works. In contrast with *grabar*, the Modern Armenian or *ašxarhabar* (”from the world”) – with two variations: Western Armenian (*arewmtahayerēn*) and Easten Armenian (*arewelahayerēn*) – arose as a literary and written language from the second half of the XIX$^{\mathrm{th}}$ century. Nevertheless, the classical language remains the foundation for specialists of the Armenian language and the key language for all philological, theological, linguistic or historic studies requiring Armenian sources prior to the XIX$^{\mathrm{th}}$ century.

Today 31.000 complete Armenian manuscripts reached us, all subsequent to the IX$^{\mathrm{th}}$ century, they only represent 5% of the total production of the Armenian copyists (Mahé, 2005 2007). Very numerous manuscripts fragments have also survived and some might constitute witnesses prior to the IX$^{\mathrm{th}}$ century, among which many ancient texts whose master copies in Latin or in Ancient Greek have been lost. Some institutions have undertaken the digitization of their collection of Armenian manuscripts or fragments, thus easing the consultation of these resources, which still requires a strong academic background in Paleography and in Classical Armenian to be read and studied. To learn and study the classical language, one must have access to precise and complete grammatical and lexicographical resources, what is sorely missing for the Armenian language, as the paper dictionaries of the XIX$^{\mathrm{th}}$ and XX$^{\mathrm{th}}$ centuries have been depleted. Although numerous initiatives exist to preserve the Classical Armenian language (digitization of manuscripts collection, online corpora and lexicographical resources), they remain limited to face the challenges of a poorly endowed language, especially regarding digital.

This article presents on one hand a situation analysis of the existing resources both printed and digital for the Classical Armenian language, and on the other hand the constitution of new resources online thanks to the Calfa platform, that has been developing multilingual complete and updated lexicographical databases for Classical Armenian, since 2014. This article aims to reflect on different matters specific to less-resourced languages, as managing the scarcity of data or building lexical databases with diverse and heterogenous written resources. Lastly, it will outline the results already achieved, the perspectives of evolution and the possible applications of these new resources. Our focus is on the process of creation and enrichment of lexical resources, assisted by crowdsourcing, and the possible integration of the database within the framework of other projects, especially regarding character recognition. Whereas it is a novelty in the Armenian Studies, the question has already been raised for other languages of the

Christian Middle East (Crane and Wulfman, 2003).

## 2. Existing Ressources for Classical Armenian

The lexicographical resources for Classical Armenian have been primarily written by the Mekhitarist Fathers of Venice, starting from the second half of the XVIII$^{\text{th}}$ century, first by realizing concordances on the basis of the Armenian Bible and the classical literature, then by creating unilingual or bilingual dictionaries of the Armenian language. Thus starting a long tradition of lexicographical erudition, to which we owe this masterpiece, unequaled to this day, the *Nor Baṙgirk‘ Haykazean Lezui* of 1836-1837 (NBHL, see *infra*). The dictionaries listed below are the main resources we have today for the Armenian studies, aside from the NBHL (non-exhaustive list):

- The *Dictionnaire abrégé français-arménien* (Concise Dictionary French-Armenian) by Father Awgerean (1812);
- The *Dictionnaire abrégé arménien-français* (Concise Dictionary Armenian-French) by Father Awgerean (1817);
- The Dictionary Armenian and English by John Brand and Father Awgerean (1821-1825);
- The *Dizionario armeno-italiano* (Dictionary Armenian-Italian) by Father J̌axǰaxean (1837);
- The Pocket Dictionary of the English Armenian and Turkish languages by Father Sōmalean (1843);
- The *Dictionnaire arménien classique-français* (Dictionary Classical Armenian-French) by Ambroise Calfa (1861);
- The *Aṙjeṙn Baṙaran Haykaznean Lezui* (Handy Dictionary of the Armenian Language) by Fathers Awgerean and Chēlalean (1846, 1865);
- The *New Dictionary Armenian-English* by Father Bedrossian (1875).

To this list, may also be added the Etymological Dictionary by Ačaṙean (Yerevan, 1926), the Explanatory Dictionary of the Armenian Language by Malxaseanc‘ (Yerevan, 1947), the Classical Armenian Dictionary of Synonyms by Łazarean (Antélias, 2006), and the Etymological Dictionary of the Armenian Inherited Lexicon by Martirosyan (Leiden, 2009).

The most significant and complete dictionary of the Armenian language is indeed the NBHL, conducted by the Fathers Awetik‘ean, Siwrmēlean and Awgerean (Venice, 1836-1837), it contains in only two volumes and 2.000 pages, all the Classical Armenian lexicon (explanatory unilingual dictionary with more than 54.000 lexical entries) as well as more than 150.000 examples from the Armenian literature, taken from the critical editions conducted by the Mekhitarist Fathers. The NBHL provides translations in numerous languages, particularly in Latin and Ancient Greek.

Most of the reliable lexicographical resources in Classical Armenian are coming from the NBHL. The resources listed above provide reduced version of the NBHL content, accessible to the general public or for conversation purposes (Calfa, 1861). Besides the limited number of lemmas and the semantic differences with the NBHL, due to the translation or the interpretation, these resources show disparities. One of the most important disparities concerns the definition and the characterization of the part-of-speech, that differ from one resource to another. The NBHL provides 14 parts-of-speech, Calfa (1861) 23 and Bedrossian 22, that results in a lack of interoperability between these databases (e.g. the lemma *aṙak‘eal* will sometimes be characterized as a nominalized past participle, sometimes as a substantive, sometimes as an adjective). Other issues of interoperability arise and have already been described in an other paper (Vidal-Gorène et al., 2019).

| | |
|---|---|
| Unilingual dictionary | 53.998 headwords |
| CA - Italian dictionary | 47.000 headwords |
| CA - English dictionary | 39.000 headwords |
| CA - French dictionary | 28.500 headwords |
| CA - Russian dictionary | 27.000 headwords |
| Synonyms dictionary | 31.472 headwords |
| Latin | 48.407 translations |
| Ancient Greek | 42.579 translations |
| Modern Armenian | 4.722 translations |
| Turkish | 2.809 translations |
| Iranian Languages | 763 translations |
| Hebrew | 701 translations |
| Arabic | 241 translations |
| Georgian | 99 translations |
| Sanskrit | 55 translations |
| Ottoman language | 23 translations |
| Syriac | 16 translations |
| Chaldean | 14 translations |
| German | 9 translations |

Table 1: Total of forms in printed resources

The table above outlines the total of headwords and translations provided in the printed resources and highlights the need for an uniform and bijective multilingual resource.

While valuable, these resources of the XIX$^{\text{th}}$ and XX$^{\text{th}}$ centuries have now been depleted. Digitized versions are available for free, in particular on the Nayiri platform that gathers 122 dictionaries in image format, work comparable to the digitization of the Latin dictionary Gaffiot, the Ancient Greek dictionary Bailly and to the Dukhrana project for Syriac (Kiraz, 2015).

Likewise, in addition to these reference dictionaries, the literary works in Classical Armenian are in need of digitization; digitized versions can already be found on platforms such as TITUS of the Johann Wolfgang Goethe University in Frankfurt (database of texts in Classical Armenian with an engine to search by form), Digilib of the American University of Armenian in Yerevan (the largest database of texts in Classical Armenian today), the eponymous platform of the Arak29 Foundation (provides particularly the Armenian Bible entirely lemmatized) and the GRE*g*ORI project of the Catholic University of Louvain (concordance of texts automatically lemmatized) (Kindt, 2018; Van Elverdinghe, 2018). The university of Leiden, under the direction of Jos Weitenberg, certainly made available, for the first time, a compilation of digital versions of texts and dictionaries (in particular the NBHL, the Ačaṙean and the Bedrossian) within the framework of the "Leiden Armenian Lexical Textbase". However, the database is no longer updated and with restricted access.

These databases are highly valuable and pionneer for the field of digitized Armenian resources. Nevertheless, they are suffering from three distinctive shortcomings, inherents to their editorial choice: firstly, most of the works and dictionaries considered are more than a century old and, besides the typos and misprints that need to be corrected, offer a content sometimes inevitably outdated and thus need to be updated; secondly, these resources, some of them not natively digital, do not enable to use and implement the latest technologies that may in fact ease and speed their consultation; lastly, it is impossible to enrich and enhance their content by offering different services and additional information. Finally, it raises the question of data interoperability, especially for texts database.
These questions have already been raised and solved for other languages of the Christian Middle East, in particular for Ancient Greek and Syriac in different projects like the Perseus Digital Library (Digital library for classical texts of the Greco-Roman world) (Smith et al., 2000), the SEDRA (dictionary, opened to crowdsourcing) of the Syriac Institute Beth Maduro, the Digital Syriac Corpus (digital library of Syriac text, opened to crowdsourcing), or the GRE*g*ORI project that offers also lemmatized resources for Ancient Greek, Syriac and Georgian (Coulie, 1996; Kindt, 2004; Kindt and Pirard, 2016; Pataridze and Kindt, 2018; Kindt, 2018), but are still quite new for Armenian. The Calfa has a similar approach in compilation, extraction, standardization and enrichment of lexical resources to the Thesaurus Linguae Graecae – dedicated to the creation and the update of corpora (Pantelia, 2000) but also proposes a lexical part – and the Ancient Greek WordNet (Bizzoni et al., 2014). The Calfa platform aims above all to offer a standardized lexical database of reference, before pursuing the constitution of massive corpora, as it is the case with these languages.

# 3. Introduction of the New Calfa Database

The lexical resource aimed by Calfa is a unified and enriched dictionary of Classical Armenian. This resource results from a compilation, comparison and unification process of the resources presented above. We outline below the essential information on this new database. For a description in-depth of this enrichment process, we would refer to (Vidal-Gorène et al., 2019).

Calfa offers, as of now, a multilingual database (Classical Armenian-French, English, Italian, and Russian) holding 65.430 homogeneous entries. For the lexicon, the dictionaries of Ambroise Calfa and Matthias Bedrossian were used as baseline. The French and the English of the original works have been updated to match the contemporary language. We proceeded to enrich and enhance the dictionaries in every language. As the NBHL is the baseline of all this work, we started, in 2016, to integrate his entire content directly to the platform and to translate all entries in the different languages already available. In the end, the platform will hold 65.430 uniformed bilingual entries. Currently, the platform holds 41.957 bilingual entries, for the rest is unilingual for now.

## 3.1. Crowdsourcing for the lexical database

The creation and the standardization of the lexical databases were made possible thanks to the implementation of collaborative work interfaces (see figure 1). A two level proofreading system was established: the general public, that is welcome to take part in the proofreading step, both with and without registration, and advanced proofreaders, selected among the Calfa team and specialists of the Armenian language. In view of the little involvement expected for a language such as Classical Armenian, we focused on the creation of a tool very easy-to-use that requires little time from the user, based on gamification targeting learners (Moirez et al., 2013). Thus, different levels were defined to enable the volunteers to choose one font over another.

1. the text is extracted with an OCR specifically trained with the fonts of the document. The Character Error Rate (CER) is below 1%.[1];
2. the raw output undergoes a double check by the volunteers and by the advanced proofreaders. The volunteers are presented with a image-text pair (at the word level) and must confirm whether the text matches the picture. If three users confirm a same pair (BnF, 2015), it is considered valid, otherwise the correction of an advanced proofreader

[1] The results achieved by the OCR on printed documents are now very good and enable the rapid constitution of corpora, as it is the case in Latin for the Corpus Corporum (Roelli, 2014) or in Syriac (Chesley et al., 2019).

| Type | Armenian | English | French | Latin | Ancient Greek | Total |
|---|---|---|---|---|---|---|
| Headwords | 54.164 | 41.975 | 32.172 | - | - | 65.430 |
| Definitions | 57.494 | 42.728 | 32.309 | - | - | 68.293 |
| Translations | 100.407 | 139.289 | 92.709 | 48.407 | 42.579 | 423.391 |
| Proper nouns | 0 | 920 | 920 | - | - | 920 |
| Synonyms | 94.635 | 94.635 | 94.635 | - | - | 94.635 |
| Etymology | 0 | 5.329 | 7.543 | - | - | 7.543 |
| Examples | 151.085 | 14.684 | 9.734 | - | - | 175.503 |

Table 2: Contents available on the Calfa platform

is required. The advanced users proofread the unprocessed data massively (either pair of words, sentences or entries) and have correction tools. Whereas the volunteers are only able to report mistakes;

3. the entries of the validated text are compared to the general database or to the most relevant dictionary already validated. This step constitutes the active phase of the resources standardization achieved by the linguists of the project (Vidal-Gorène et al., 2019), and the correction of the mistakes of the original document. If necessary, these data are used to train the OCR again;

4. the data are uploaded online after the update of the vocabulary (e.g. outdated meaning in French) if case needed.

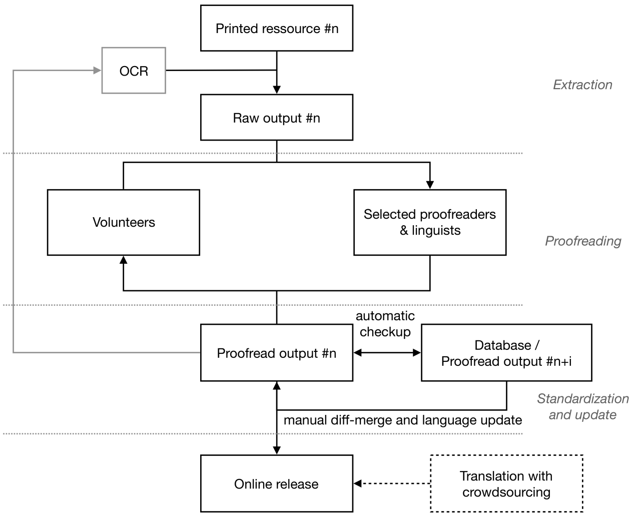

Printed ressource #n
OCR
Extraction
Raw output #n
Volunteers
Selected proofreaders & linguists
Proofreading
automatic checkup
Proofread output #n
Database / Proofread output #n+i
Standardization and update
manual diff-merge and language update
Online release
Translation with crowdsourcing

Figure 1: Creation and Evaluation Process of the databases

The main works happened between 2014 and 2016. Although eighty people registered for the correction process, merely a quarter was truly active, due to the lack of short term goals.[2] The median correction of volunteers was 1.618 words a month. The same interface was implemented for the correction of the HTR results (see *infra* Database Integration and Text Recognition).

To process the problematic cases in Latin and in Ancient Greek, our linguists compared the NBHL data with the dictionaries of Classical Armenian-Greek by Čōlagean (Constantinople, 1868) and Classical Armenian-Latin by Misk‘čean (Rome, 1887). Then the forms in Latin and Ancient Greek have first been verified and analyzed by Collatinus and Eulexis respectively (Ouvrard and Verkerk, 2017), and then by GRE*g*ORI. The 150.000 examples from the NBHL, as well as the Armenian Bible, have been automatically lemmatized thanks to morphological engines (in cooperation with the GRE*g*ORI project), see *infra*, and corrected manually through a crowdsourcing interface (Clérice et al., 2019). Unlike the Perseus Digital Library that calls on users to suggest analysis and to vote for the most accurate analysis to be posted, we favor an automatic analysis seconded by the correction of users trained beforehand through instructional games. An ongoing project with the University of Louvain is comparing the results of the lemmatization process. These examples will be translated later on. The plaform offers also a dictionary of synonyms based on the work by Łazarean and an etymological dictionary extracted from the Etymological Dictionary of the Armenian Language by Ačaṙean, whose translation in French and in English we started. In addition, we put online a dictionary for proper nouns (surnames and toponyms) based on the proper nouns collated by J̌aǰaxean (1837) and Awgerean in his condensed dictionary Armenian-French (Venice, 1812).
On completion, the Calfa platform will provide:

- a Classical-Armenian-French Dictionary;
- a Classical-Armenian-English Dictionary;
- a Classical-Armenian-Italian Dictionary;
- a Classical-Armenian-Russian Dictionary;
- a unilingual Classical-Armenian Dictionary;
- a Classical-Armenian-Ancient Greek Dictionary;
- a Classical-Armenian-Latin Dictionary;
- a Synonyms Dictionary;

[2] This limit was evidenced in 2016, and has since been corrected for the creation and implementation of new tools, that will be described later on.

- an Etymological Dictionary;
- a Proper Nouns Dictionary.

The table 2 below outlines the distribution of the data collated and processed by Calfa and currently available on the platform. The process of data enrichment and standardization is still ongoing. The goal is to reach the same total of lemmas for each language. All in all, 65.430 entries have been identified and comprise the data presented in table 2.

### 3.2. Morphological Additions

Due to the evident shortage of space, inherent of the paper medium, a dictionary never contains every possible form one entry might have. In Classical Armenian, the declension holds seven cases (nominative, accusative, genitive, dative, ablative, locative and instrumental), in singular and plural. The conjugation is based on two verbal stems (present and past), has two voices (active and medio-passive), four finite verbs (indicative, subjunctive, imperative and exhortative), two non-finite verbs (infinitive and participle) and three tenses (present, preterit and aorist). After studying the noun-forms and verb-forms collated in the texts, we worked on inflectional morphology to develop morphological engines that conjugate and inflect automatically every entry, and generate around 2.000.000 forms. All in all, our inflection table holds up to 76 cells for the verbs depending on the conjugation and up to 14 cells for the declension.
However, the structure and display of morphosyntactic informations is different for each dictionary, we generalized the following minimum structure[3] to characterize quickly the inflection types and to achieve the declension and conjugation automatically:

- canonical form, inflectional ending (genitive singular and genitive(s) plural);
- verbal form present first person singular, aorist first person singular. For the verbs, we defined 19 groups in order to cover all the regular forms.

| Field | Example |
|---|---|
| Entry of dictionary | *azatanam, ec'ay* |
| Lemma | *azatanam* |
| Inflectional ending selected | *-anam* |
| Morphological complement | *ec'ay* |
| Infinitive inflectional ending | *-anal* |
| Inflectional ending (36 cells) | *-anam, -anas, -anay, etc.* |

Table 3: Example of the structure of the verb inflection table

The inflectional ending selected doesn't necessarily match the morphological segmentation of the lemma, but has been chosen in order to embrace the greatest number of entries unequivocally. The inflectional ending of conjugated forms is added to the root to create the inflected forms. In case of a defective form or in case the form is not attested either in Calfa or in other databases, the inflectional ending is replaced by a dash. If, for one cell, several forms are possible, they are included in the cell (and separated by pipes in code). The irregular forms have been typed manually.

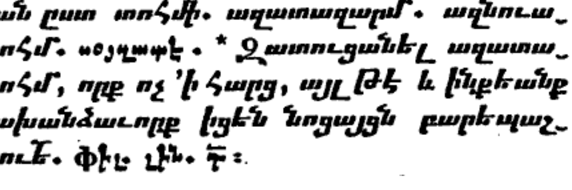
Such a development doesn't present particular difficulties, because of the thorough knowledge of the irregular entries, of the predictable vocalic alternation and of the identification of the monosyllabic entry governed by specific rules (e.g. augment aorist third person singular). Thus designed, the system enables to generate more than 1.000.000 forms. Hence, combined with the forms attested in the corpus, Calfa offers close to 2.000.000 forms (including homographies, apposed prefix, etc.). In the end, other informations will be added to this database. The morpho-lexical tags used in the texts follow the rules laid out for Classical Armenian by the collaborators of the GRE*g*ORI project (Coulie et al., 2020).

### 3.3. Display of The Results

Among the ten dictionaries previously mentioned and available on Calfa only four are to be considered as leading in the constitution of the platform, because they have their own separate interface: the French, the English, the Italian and the Russian dictionaries. The dictionary of proper nouns is also considered as a separated independent database. The NBHL, the Latin, the Ancient Greek, the Etymological and the Synonyms dictionaries content are directly integrated to the entries of the four others.

## 4. Database Integration and Text Recognition

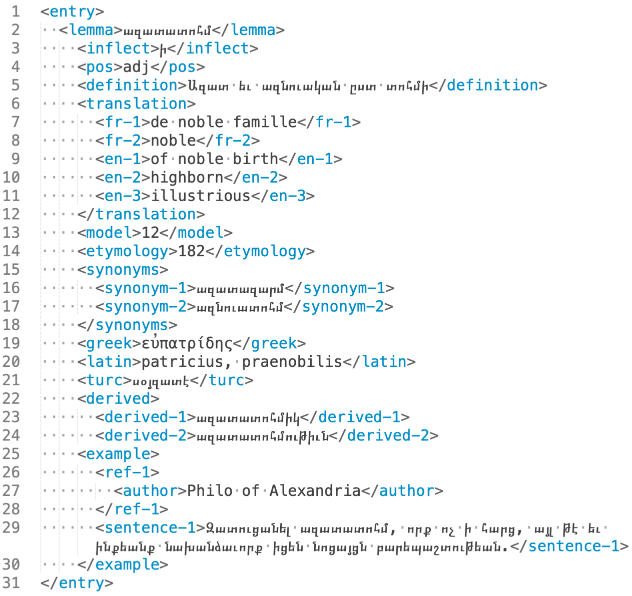
The database briefly described here is already used by different projects. All the entries are linked thanks to the digital medium and the lemmatization. This results generation benefits from every dictionary within the platform and offers to the user an extensive overview of the Classical Armenian language.



ԱԶԱՏԱՏՈՀՄ, ի. ա. Εὐπατρίδης. Patricius, praenobilis. Ազատ և ազնուական ըստ տոհմի. ազատազարմ. ազնուատոհմ. աօյղափէ. * Զատուցանել ազատատոհմ, որք ոչ 'ի հարց, այլ թէ և ինքեանք նախանձաւորք լիցեն նոցայցն բարեպաշտուէ. Փիլ. լին. ՟Գ:

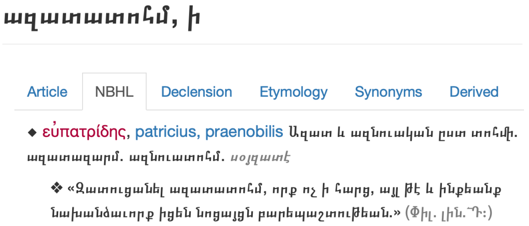
Figure 2: Headword *azatatohm* from NBHL (I.27)

[3] For irregular entries, we manually add the inflectional endings or the complementary forms.



```
<entry>
  <lemma>ազատատոհմ</lemma>
    <inflect>ի</inflect>
    <pos>adj</pos>
    <definition>Ազատ եւ ազնուական ըստ տոհմի</definition>
    <translation>
      <fr-1>de noble famille</fr-1>
      <fr-2>noble</fr-2>
      <en-1>of noble birth</en-1>
      <en-2>highborn</en-2>
      <en-3>illustrious</en-3>
    </translation>
    <model>12</model>
    <etymology>182</etymology>
    <synonyms>
      <synonym-1>ազատազարմ</synonym-1>
      <synonym-2>ազնուատոհմ</synonym-2>
    </synonyms>
    <greek>εὐπατρίδης</greek>
    <latin>patricius, praenobilis</latin>
    <turc>սօյզատէ</turc>
    <derived>
      <derived-1>ազատատոհմիկ</derived-1>
      <derived-2>ազատատոհմութիւն</derived-2>
    <example>
      <ref-1>
        <author>Philo of Alexandria</author>
      </ref-1>
      <sentence-1>Զատուցանել ազատատոհմ, որք ոչ ի հարց, այլ թէ եւ
      ինքեանք նախանձաւորք իցեն նոցայցն բարեպաշտութեան.</sentence-1>
    </example>
</entry>
```

Figure 3: Example of the informations compiled in the Calfa database for the headword *azatatohm*, before lemmatization



**ազատատոհմ, ի**

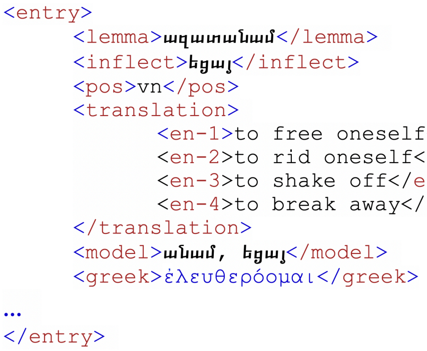
◆ εὐπατρίδης, patricius, praenobilis Ազատ և ազնուական ըստ տոհմի. ազատազարմ. ազնուատոհմ. սօյզատէ

❖ «Զատուցանել ազատատոհմ, որք ոչ ի հարց, այլ թէ և ինքեանք նախանձաւորք իցեն նոցայցն բարեպաշտութեան.» (Փիլ. լին.՟Դ:)

Figure 4: Display of the headword *azatatohm* on the Calfa website - all the words are links

All the search functionalities integrated to the platform (advanced search, word construction layout, concordance layout, full-text search, lemmatized corpus, search by keywords, by closest form, by theme, by inflected form, by grammatical category, etc.) have been made possible by the process of standardization of all entries and the development of morphological engines. These functionalities will be described in an other paper.

We implemented an API that enables to query the Calfa database. Data are structured as XML files or JSON files, where are featured the lemmas along with all content described so far (parts-of-speech, generated forms, definitions, homonyms, synonyms, etc.). This API can be directly integrated to external corpus, as in the case of the GRE*g*ORI corpus, or be used for form generation in the framework of Natural Language Processing particularly, or for morphosyntactic labels suggestion (lacking of graphical standardization today), etc. The informations returned by the API may concern all levels of the data tree for a given entry, provided the entry is in the database.



```
<entry>
      <lemma>ազատանամ</lemma>
      <inflect>եցայ</inflect>
      <pos>vn</pos>
      <translation>
            <en-1>to free oneself</en-1>
            <en-2>to rid oneself</en-2>
            <en-3>to shake off</en-3>
            <en-4>to break away</en-4>
      </translation>
      <model>անամ, եցայ</model>
      <greek>ἐλευθερόομαι</greek>

...
</entry>
```

Figure 5: Example of a possible structure provided by the API

Calfa conducts also researches in automatic recognition of the Armenian handwrittings (Handwritten Text Recognition), for ancient manuscripts in particular, in the framework of the Vision Calfa Project. We use the database for the post-processing of the HTR results, thanks to the data labeled in context (n-gram, BERT, etc.) (Rigaud et al., 2019). This is the main application of the database today. In return, the recognized handwritten data feed the text database, via crowdsourcing interfaces.

Faced with the shortage of data and the limited number of trained users, we have developed solutions and crowdsourcing tools to enable both the gradual improvement of the HTR and the growth of an involved and effective community regardless to the individual's knowledge of the language.

Indeed, in order to train the HTR technology, we developed and implemented two crowdsourcing tools for medieval Armenian manuscripts labeling and for the correction of the systems predictions. These solutions have been customized for a non-specialist public, and do not need to be an Armenian-speaking public. This approach is critical for a poorly endowed language, whose specialists are becoming scarce.

(i) The first interface proposes to identify lines of text within a page of a manuscript, and to type, when the prediction is incorrect, the corresponding ground truth.

(ii) The second tool is a validation technique for the system predictions, that enables us to inspect the database: each data (character, word or line) needs to be read by three users, as we did for the lexical database proofreading (BnF, 2015), see *supra*. This step helps improve the system by eliminating the noise linked to the automatic processing. It is all the more important because it enables to generate large quantity of data and to refine the HTR results. We are using neural networks, and combine indeed both word-based and line-based learning. The Calfa platform offers now more than 150.000 handwritten characters, 29.000 words and 11.000 lines for the four Armenian scripts. These data have been generated by 122 volunteers.

The Vision Calfa database, that will be described later on, is the very first for the Armenian language, which is largely under-resourced regarding such systems. The corpus generated from the study of ancient manuscripts are then added to the text database of Calfa in order to increase the number of annotated examples on the platform.

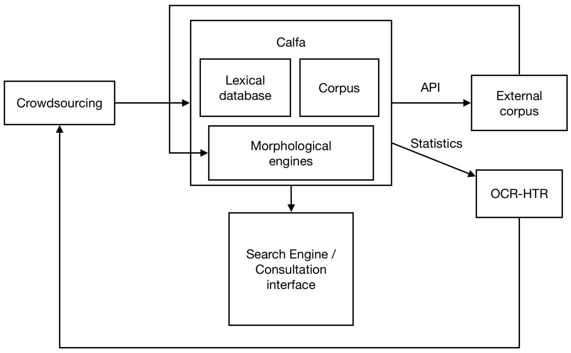


Figure 6: Process and of use of the Calfa

## 5. Conclusion

Classical Armenian combines the characteristics specific to the poorly endowed ancient languages and the challenges of preservation and understanding of living languages. Faced with the difficulties to access and use the data, the project described in this article offers innovative solutions in tune with the reality of the language and its users. Coping with limited scattered and heterogeneous resources, the construction of a reference linguistic platform (multilingual dictionaries, morphological engines, texts databases, etc.) is the first step of the process of language accessibility and preservation. Faced with the difficulties to use the Classical Armenian sources, this project combines precise paleographical studies and latest innovations in computer vision to offer a relevant and efficient HTR software.

Calfa is the richest database for Classical Armenian today, both in lemmas and translations, and in generated and attested forms. To that end, we focused on crowdsourcing. The structure of data enables its integration in other projects, in particular projects of NLP via its API and is designed to enrich and build corpora for Classical Armenian. The contents are currently being interfaced on the Calfa platform, that provides a very complete search engine, dedicated to the specific expectations and needs of the researchers, but also to the discovery of the Armenian lexicon by beginners. The etymological contents in Italian and Russian should be available in 2020. The lemmatized data, as well as the attested and not attested, are now representative of the classical language state. The approach of lemmatization through rule-dictionary, adequate for the corpus currently processed, has obvious limitations and is inadequate to process massively and quickly various corpora with new medieval language state (Van Elverdinghe, 2018; Kindt, 2018). We are experimenting to realize lemmatization by using joint learning (with the previously annotated corpus), notably with the view to analyze unknown forms and to reduce manual correction time. Armenian and especially Classical Armenian is a language still poorly endowed although, paradoxically, there are numerous digitization and corpora projects. The purpose of the Calfa platform is to offer the first interoperable database, both for linguistic resources and for lines and handwritten characters' databases, in order to integrate them within larger systems and developments for the Armenian language.

## 6. Acknowledgement

The Calfa platform has been funded in 2014 by the Calouste Gulbenkian Foundation and the Brothers Ghoukassiantz Foundation, and in 2016 by the Calouste Gulbenkian Foundation in the context of the call for project ”Armenian Language in the Digital Age”. We especially thank our volunteers for their help in the proofreading process.

## 7. Bibliographical References

Awetikʻean, G., Siwrmēlean, X., and Awgerean, M. (1837–1837). *New Dictionary of the Armenian Language*. Tparan i Srboyn Łazaru, Venice.

Ačaṙean, H. (1926). *Etymological Dictionary of the Armenian Language*. EPH, Yerevan.

Bedrossian, M. (1875). *New Dictionary Armenian-English*. Tparan i Srboyn Łazaru, Venice.

Bizzoni, Y., Boschetti, F., Diakoff, H., Gratta, R. D., Monachini, M., and Crane, G. (2014). The making of ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1140–1147. European Language Resources Association (ELRA).

BnF. (2015). Réalisation d'une étude d'usages des utilisateurs de la plateforme experimentale correct. Technical Report ACM 248, Bibliothèque Nationale de France.

Calfa, A. (1861). *Dictionnaire arménien-français*. L. Hachette et cie, Paris.

Chesley, E., Marcantonio, J., and Pearson, A. (2019). Towards digital syriac corpora: Evaluation of tesseract 4.0 for syriac ocr. *Journal of Syriac Studies*, 22(1):109–192.

Clérice, T., Pilla, J., Camps, J.-B., and architexte. (2019). https://github.com/hipster-philology/pyrrha: 2.1.0.

Coulie, B., Kindt, B., and Kepeklian, G. (2020). Étiquettes morphosyntaxiques et flexionnelles pour le traitement automatique de l'arménien ancien. *Études Arméniennes Contemporaines*. in press.

Coulie, B. (1996). La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs. *Byzantion*, 66:35–54.

Crane, G. and Wulfman, C. (2003). Towards a cultural heritage digital library. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings.*, pages 75–86.

Donabédian, A. (1994). Quelques remarques sur l'alphabet arménien. *Slovo*, 14:7–21.

Ghazarean, R. (2006). *Classical Armenian - Dictionary of Synonyms.* Armenian Catholicosate of Cilicia, Antelias.

Hübschmann, H. (1897). *Armenische Grammatik.* Von Breitkopf & Härtel, Leipzig.

J̌axǰaxean, M. (1837). *Dizionario armeno-italiano.* Tparan i Srboyn Łazaru, Venice.

Kindt, B. and Pirard, M., (2016). *De Nazianze à Ninive. La couverture lexicale du Dictionnaire Automatique Grec*, pages 49–77. Orientalia Lovaniensia Analecta 25. Peeters, Leuven.

Kindt, B. (2004). La lemmatisation des sources patristiques et byzantines au service d'une description lexicale du grec ancien : les principes de formulation des lemmes du dictionnaire automatique grec. *Byzantion*, 74:213–272.

Kindt, B. (2018). Processing tools for Greek and Other Languages of the Christian Middle East. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages.

Kiraz, G. (2015). Automatic splitting of lexical entries from classical dictionaries. *Pacific journal*, 10.

Lamberterie, C. (1992). Introduction à l'arménien classique. *Lalies*, 10:233–289.

Mahé, J.-P. (2005-2007). Une approche récente de la paléographie arménienne. *Revue des études arméniennes*, 30:433–438.

Makarean, A. (1838). *Armenian-Russian Dictionary.* I tpagratan čemarani Tearc' Lazareanc', Moscow.

Martirosyan, H. (2009). *Etymological Dictionary of the Armenian Inherited Lexicon.* Brill, Leiden.

Moirez, P., Moreux, J.-P., and Josse, I. (2013). État de l'art en matiere de crowdsourcing dans les bibliotheques numeriques. Technical Report L-4.3.1, Bibliothèque Nationale de France.

Ouvrard, Y. and Verkerk, P. (2017). Collatinus & eulexis: Latin & greek dictionaries in the digital ages. In *Digital Classics III: Re-thinking Text Analysis.* Center for Hellenic Studies/Harvard University.

Pantelia, M. (2000). 'Noûs, into CHAOS': The Creation of the Thesaurus of the Greek Language. *International Journal of Lexicography*, 13(1):1–11.

Pataridze, T. and Kindt, B. (2018). Text alignment in Ancient Greek and Georgian: A case-study on the first homily of gregory of nazianzus. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages.

Rigaud, C., Doucet, A., Coustaty, M., and Moreux, J.-P. (2019). ICDAR 2019 Competition on Post-OCR Text Correction. In *15th International Conference on Document Analysis and Recognition.* ICDAR.

Roelli, P. (2014). The corpus corporum, a new open latin text repository and tool. *ALMA (Archivum Latinitatis Medii Aevi)*, 78.

Smith, D. A., Rydberg-Cox, J. A., and Crane, G. (2000). The Perseus Project: a digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Van Elverdinghe, E. (2018). Recurrent Pattern Modelling in a Corpus of Armenian Manuscript Colophons. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages.

Vidal-Gorène, C., Decours-Perez, A., Queuche, B., Ouzounian, A., and Riccioli, T. (2019). Digitalization and enrichment of the nor baṙgirk' haykazean lezui: Work in progress for armenian lexicography. *Journal of the Society of Armenian Studies*, 27. in press.

## 8. Language Resource References

American University of Armenia. (1999). *Digital Library of Armenian Literature.*

Arak29. (2002). *Arak29.*

Calfa. (2014). *Calfa - Enriched Dictionaries of Classical and Modern Armenian.*

G. Crane. (1985). *Perseus Digital Library.* Tufts University.

J. Gippert. (2003). *TITUS Project.* Johann Wolfgang Goethe University.

G. Kiraz. (1988). *Syriac Electronic Data Research Archive (SEDRA).* Beth Mardutho - The Syriac Institute.

L. J. Lindgren. (2006). *Dukhrana.* Dukhrana Biblical Research.

Nayiri. (2004). *Nayiri platform.*

(2014). *Thesaurus Linguae Graecae.* University of California.

Université Catholique de Louvain. (1990). *GREgORI Project - Softwares, linguistic data and tagged corpus for ancient GREek and ORIental languages.*

J. E. Walters. (2004). *Digital Syriac Corpus.*

J. Weitenberg. (2000). *The Leiden Armenian Lexical Textbase.* University of Leiden.