NEWS EYE

NewsEye International Conference
March, 17th 2021

# Two Examples of Analysis of Textual Document in Oriental and Under-Resourced Languages

**Chahan Vidal-Gorène**
Calfa, École Nationale des Chartes-PSL

CALFA



**Open databases** for public and NLP Projects

**Document Analysis** for Under-Resourced Languages

**Customized IT Projects** for Libraries

Specialized in **Automatic Document Analysis**

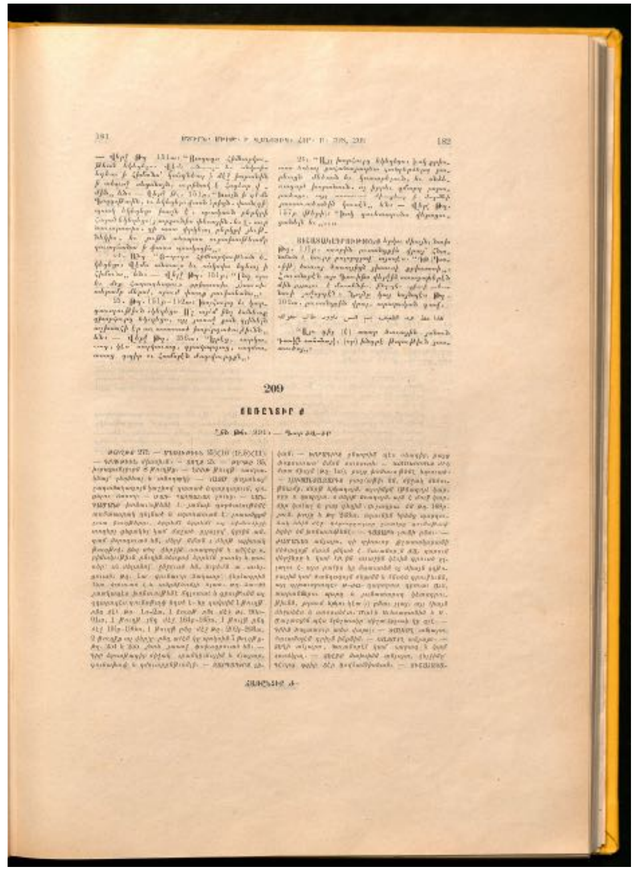
➡ **Handwritten** and Printed **Text Recognition** (HTR / OCR) for Oriental Languages and **structure understanding**

# Ongoing projects for Armenian Newspapers


https://webaram.com/biblio/presse/haratch-յառաջ

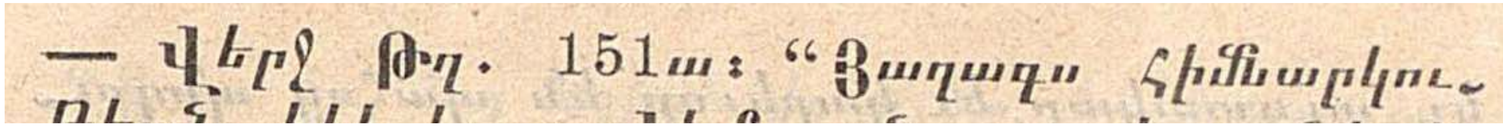
Digitized by BULAC and ARAM association
with the support of the Gulbenkian Foundation

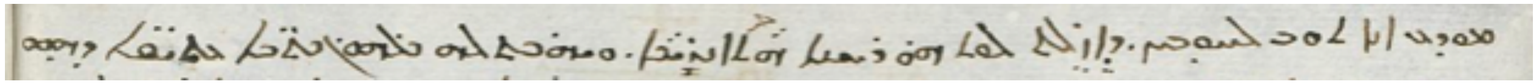

https://arar.sci.am/dlibra/collectiondescription/10
Digitized by the Fundamental Scientific Library
of the NAS RA

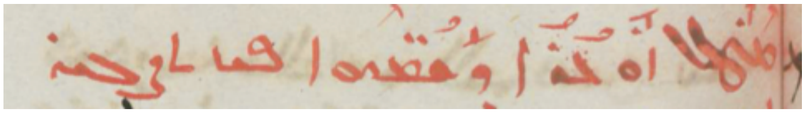


Digitized by Calfa and Mekhitarist
Congregation

**Previously OCRized, with noise in Layout extraction and CER ~ 85% (Tesseract)**







**Issues** with Under-Resourced Languages (or very specific layouts)

**Need** of Annotated Databases

but for which kind of project ?
OCR ? NER ? etc

**Lack** of interoperability of current models

**Lack** of specialists to annotate manually

**Help of Fine Tuning ?**

**Crowdsourcing ?**

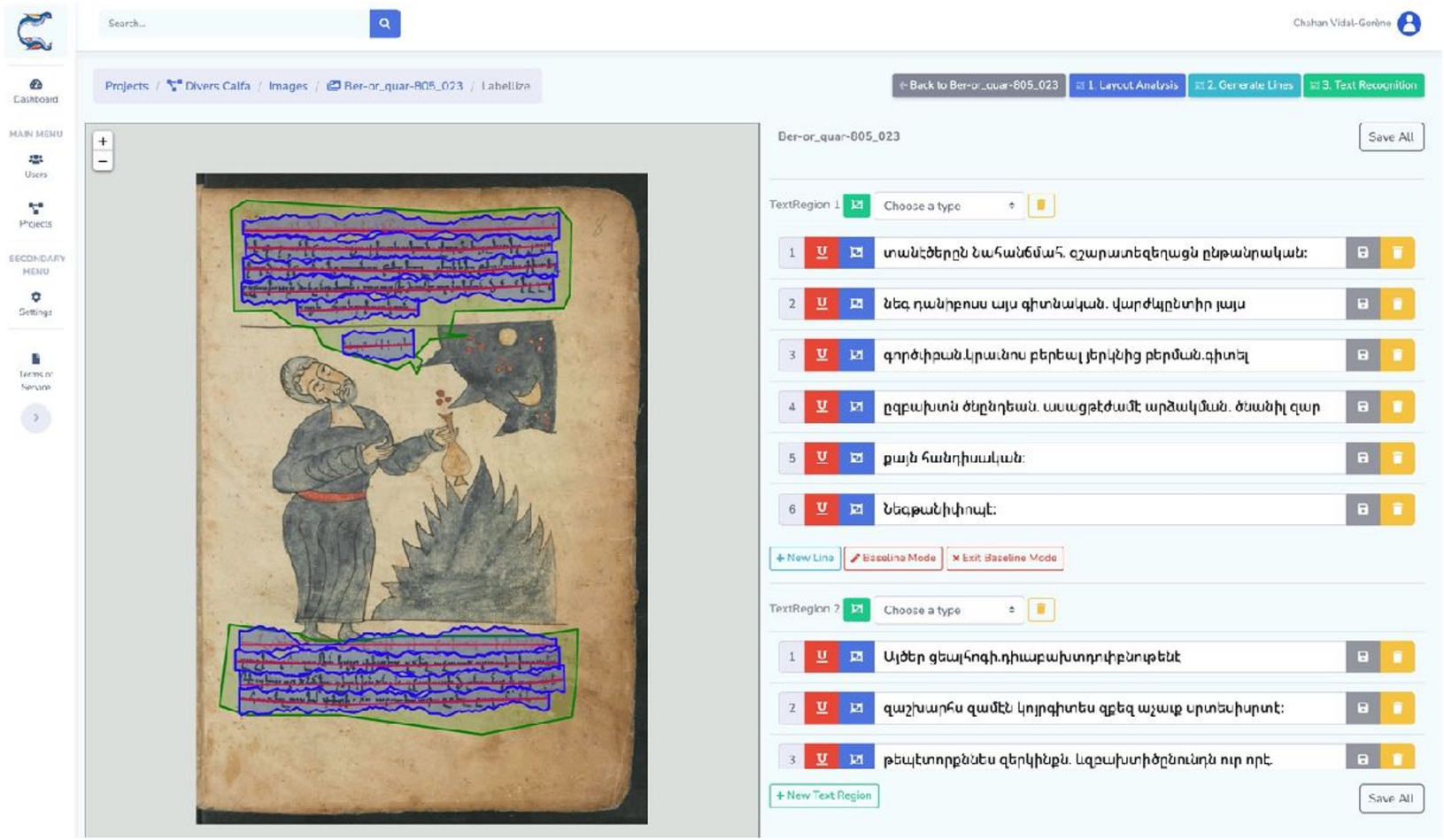
of which technology ?

of which models ?

with which expertise ?

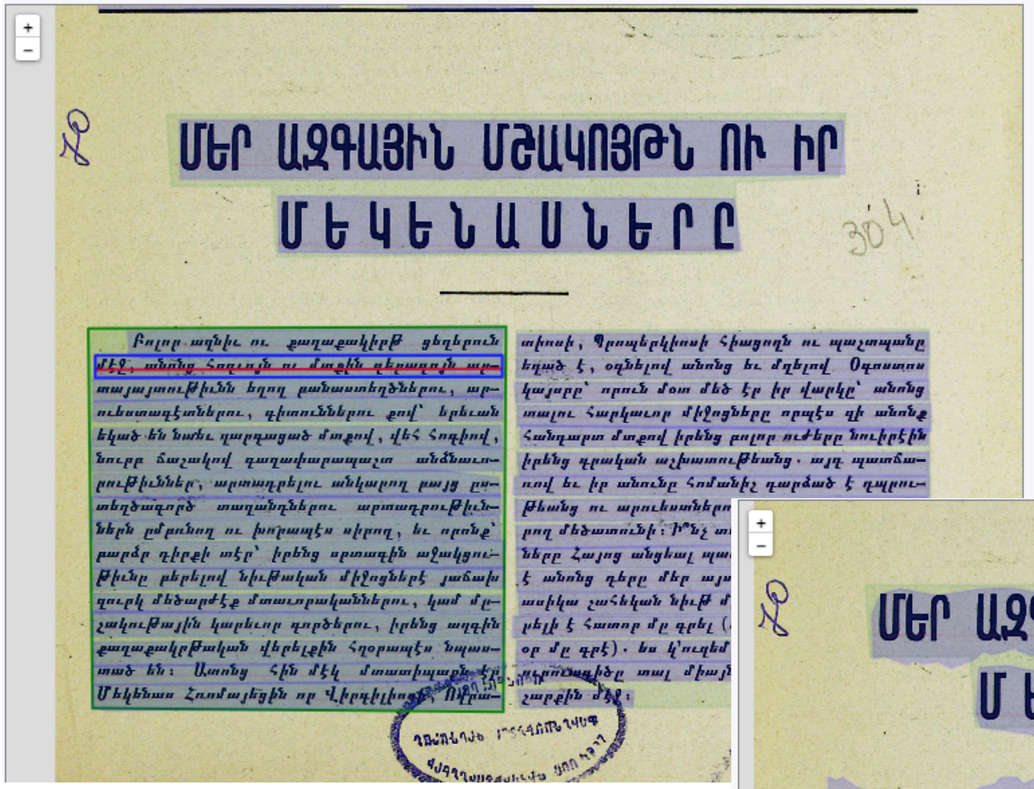
**https://calfa.fr/vision**

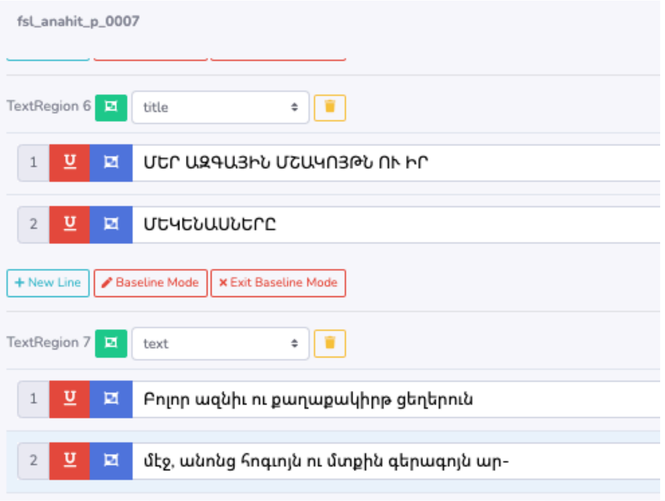
Calfa Vision is a web-based annotation tool for documents and images, collaborative and free



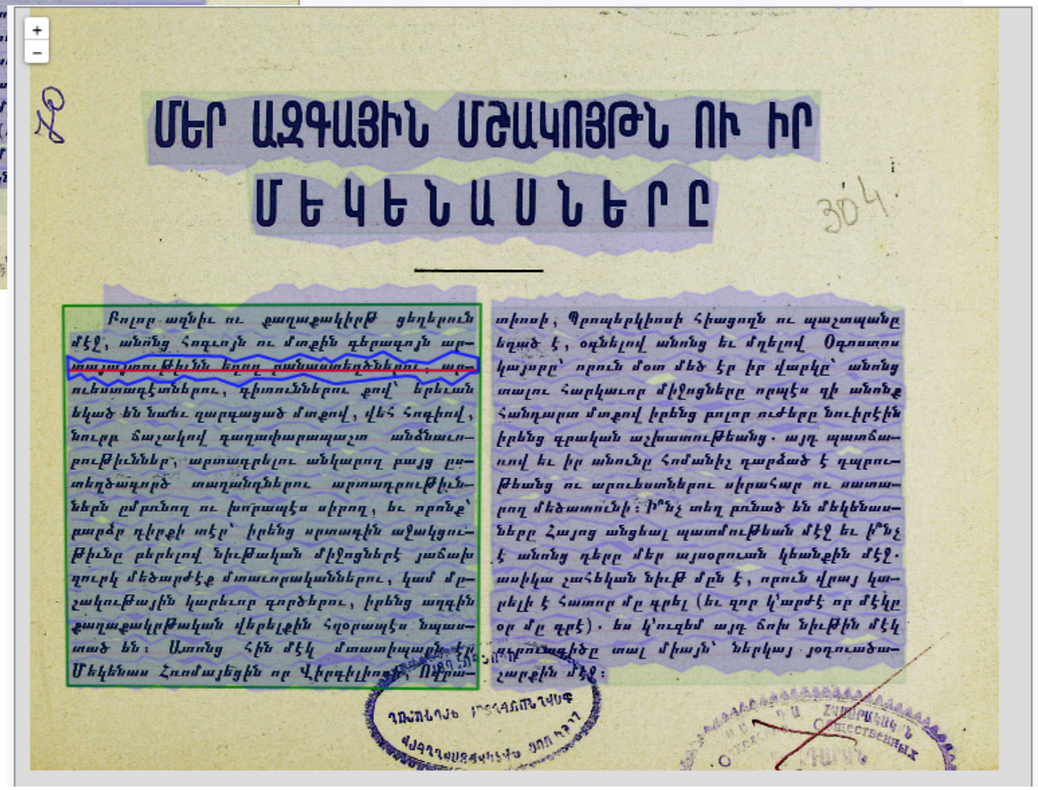

- Originally developed for Historical Manuscripts in Oriental Languages
- Includes automated tasks and fine-tunable models in real time

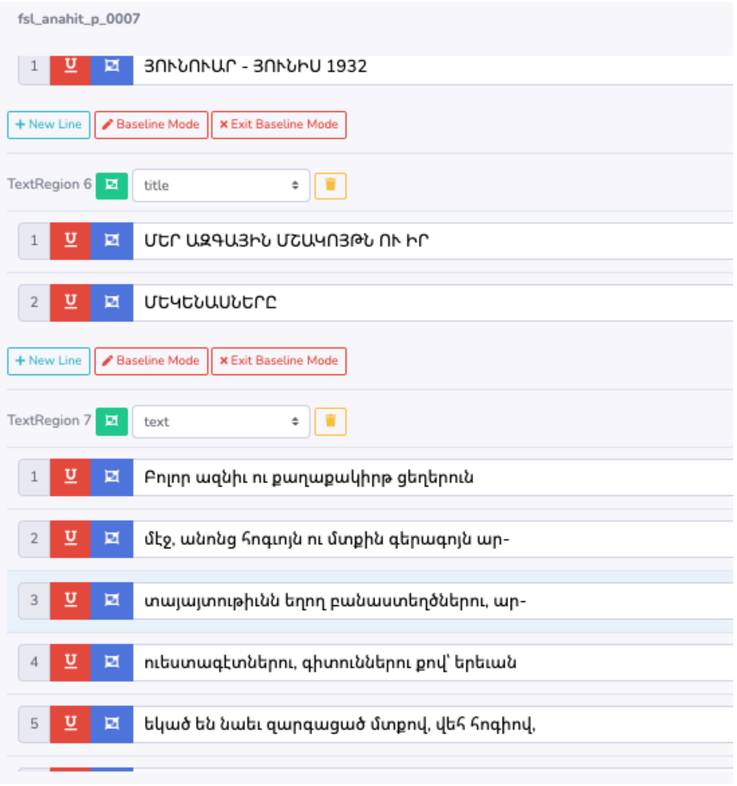
=> applied to process newspapers

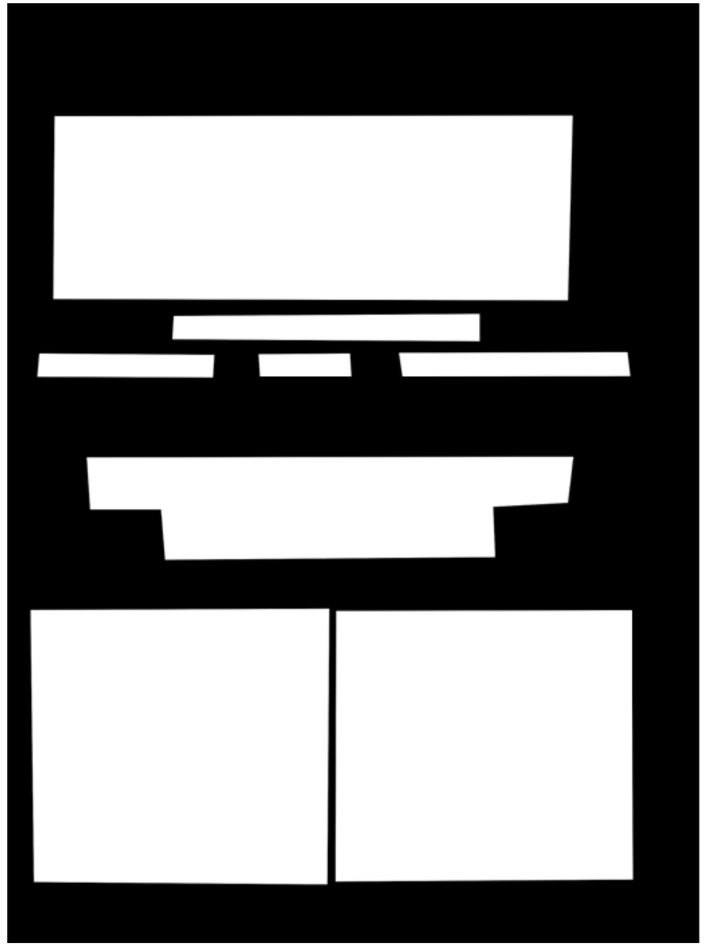


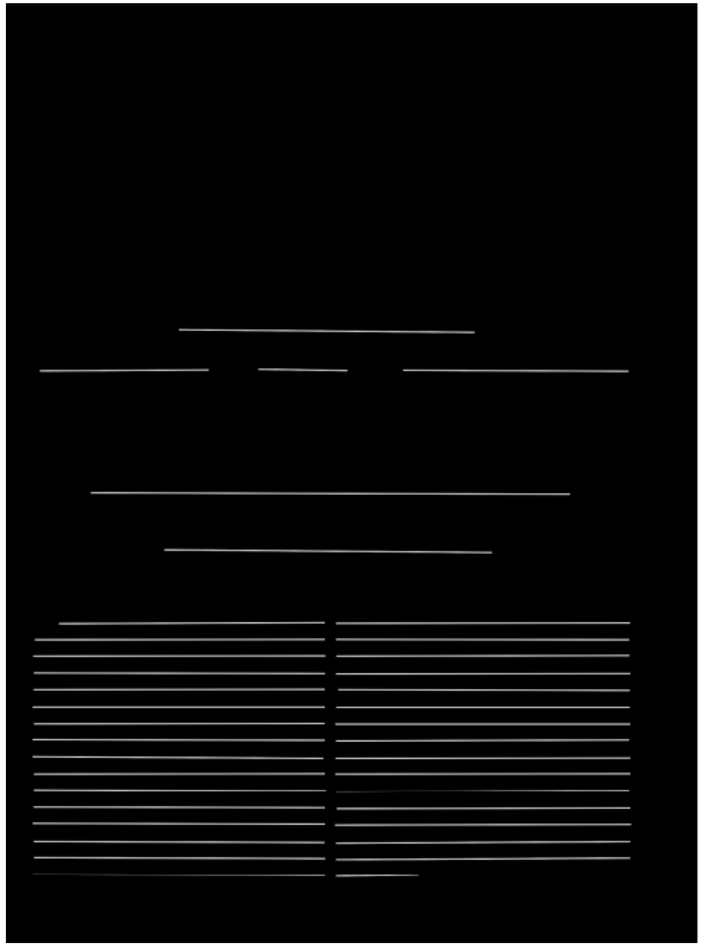




Automatic analysis (layout analysis and OCR) of Anahit Journal, Fundamental Scientific Library (Armenia)



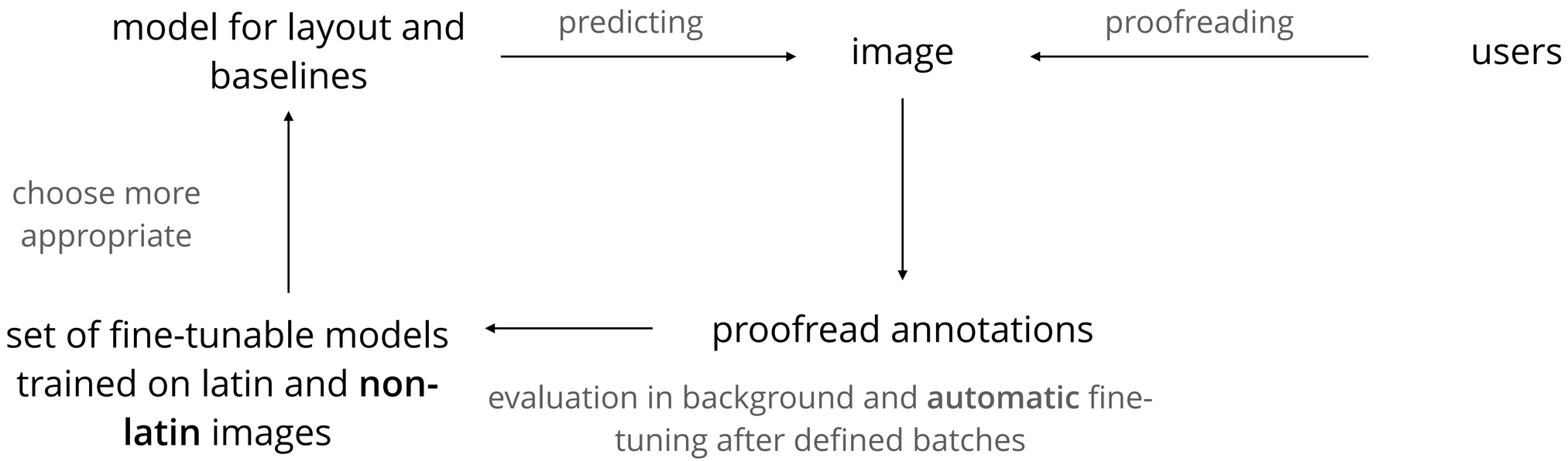




```
<?xml version="1.0" ?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013
  <Metadata>
    <Creator>Calfa</Creator>
    <Created>2021-03-17T10:33:01+00:00</Created>
  </Metadata>
  <Page imageFilename="fsl_anahit_p_0007.jpg" imageHeight="2938" imageW
    <TextRegion id="8398" custom="structure {type:heading;}">
      <Coords points="149,353 1766,351 1752,927 145,923 149,353"/>
      <TextEquiv>
        <Unicode></Unicode>
      </TextEquiv>
    </TextRegion>
    <TextRegion id="8399" custom="structure {type:title;}">
      <Coords points="521,977 1476,971 1476,1053 517,1047 521,977"/>
      <TextLine id="99443">
        <Coords points="539,965 576,961 620,971 683,964 709,951 744,972
        <Baseline points="539,1020 1468,1028"/>
        <TextEquiv>
          <Unicode>ՀԱՆԴԷՍ ՄՏԱԾՄԱՆ ԵՒ ԱՐՈՒԵՍՏԻ</Unicode>
        </TextEquiv>
      </TextLine>
      <TextEquiv>
        <Unicode>ՀԱՆԴԷՍ ՄՏԱԾՄԱՆ ԵՒ ԱՐՈՒԵՍՏԻ</Unicode>
      </TextEquiv>
    </TextRegion>
    <TextRegion id="8400" custom="structure {type:numerotation;}">
      <Coords points="99,1097 643,1101 639,1169 93,1167 99,1097"/>
      <TextLine id="99444">
        <Coords points="104,1104 157,1091 194,1094 207,1104 231,1095 20
        <Baseline points="104,1147 632,1144"/>
        <TextEquiv>
          <Unicode>Գ. ՏԱՐԻ, ԹԻՒ 5–6</Unicode>
        </TextEquiv>
      </TextLine>
      <TextEquiv>
        <Unicode>Գ. ՏԱՐԻ, ԹԻՒ 5–6</Unicode>
      </TextEquiv>
    </TextRegion>
```

**Output** : text, ALTO, pageXML, customized json, masks, pairs GT

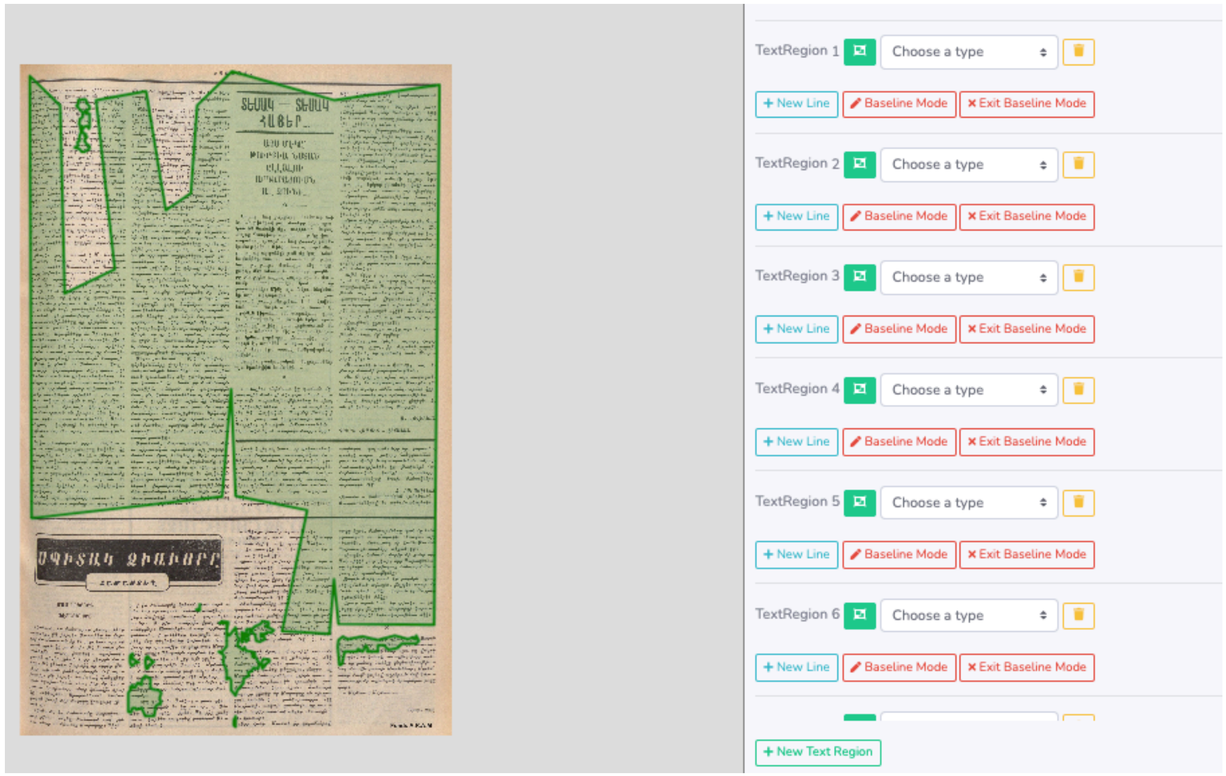
# Pipeline to reduce OCR/HTR correction and proofreading




Basically, we could also considered that damaged documents consist in themself a under-resourced variation.

| Dataset | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| BADAM | 0.8253 | 0.8541 | 0.8395 |
| cBAD complex track (ICDAR 2017) | 0.9071 | 0.9052 | 0.9061 |
| cBAD simple track (ICDAR 2017) | 0.9511 | 0.9538 | 0.9525 |
| cBAD (ICDAR 2019) | 0.9312 | 0.9310 | 0.9311 |
| READ (ICFHR 2016) | 0.9590 | 0.9888 | 0.9737 |
| Training Dataset | 0.9531 | 0.9878 | 0.9701 |

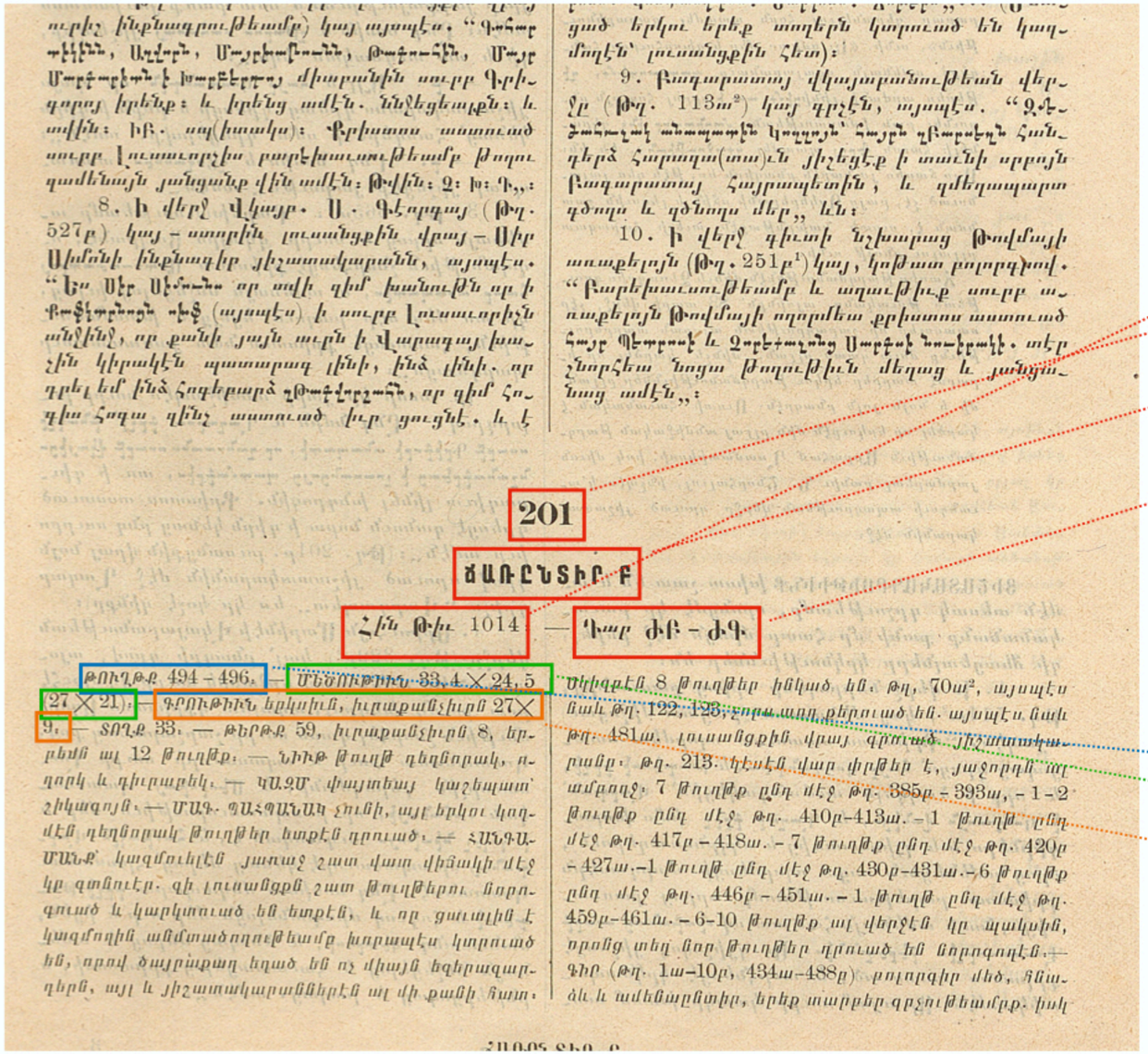
Vidal-Gorène et al, 2021, *submitted*




default model, 1st image


fine-tuned model (batch 20)

On average, text class pixel accuracy 87,03 % - results of fine-tuning on Calfa Vision after 20 pages manually proofread



201

ՃԱՌԸՆՏԻՐ Բ

Հին թիւ 1014 — Վար ԺԲ – ԺԳ

Թուղթք 494 – 496. — Մեծութիւն 33,4×24,5 (27×21). — Գրութիւն երկսիւն, իւրաքանչիւրն 27×9. — Տողք 33. — Թերթք 59, իւրաքանչիւրն 8, երբեմն ալ 12 թուղթք. — Նիւթ թուղթ դեղնորակ, ոզորկ և փխրաբեկ. — Կազմ փայտեայ կաշեպատ՝ շիկագոյն. — Մաշ. պահպանած չունի, այլ երկու կողմէն դեղնորակ թուղթեր եռաքէն դրուած. — Անդամանք կազմուելէն յառաջ շատ վատ վիճակի մէջ կը գտնուէր. զի լուսանցքն շատ թուղթերու ծորգուած և կարկտուած են եռքէն, և որ ցաւալին է կազմողին անմտածողութեամբ խորապէս կտրուած են, որով ծայրաքաղ եղած են ոչ միայն եզերազարդերն, այլ և յիշատակարանիներէն ալ մի քանի հատ.

Սկիզբն 8 թուղթեր ինկած են, թղ. 70ա², այսպէս նաև թղ. 122, 123, ուրկէ աղէ քերուած են. այսպէս նաև թղ. 481ա. Լուսանցքին վրայ գրուած յիշատակարանք. թղ. 213. Դէպի վար փրթեր է, յաջորդն ալ ամբողջ. 7 թուղթք ընդ մէջ թղ. 385բ – 393ա, – 1 – 2 թուղթք ընդ մէջ թղ. 410բ–413ա. – 1 թուղթ ընդ մէջ թղ. 417բ – 418ա. – 7 թուղթք ընդ մէջ թղ. 420բ – 427ա. – 1 թուղթ ընդ մէջ թղ. 430բ–431ա. – 6 թուղթք ընդ մէջ թղ. 446բ – 451ա. – 1 թուղթ ընդ մէջ թղ. 459բ–461ա. – 6–10 թուղթք ալ վերջէն կը պակսին, որոնց տեղ նոր թուղթեր դրուած են նորագոյնս. — Գիր (թղ. 1ա–10բ, 434ա–488բ) բոլորգիր մեծ, հնաձև և ամենաընթեր, երեք տարբեր գրչութեամբք. իսկ

IDENTIFIANT V1014

NUMÉRO 1014

NUMÉRO DE NOTICE 201

TITRE ՃԱՌԸՆՏԻՐ Բ

DATE DE DÉBUT XII

DATE DE FIN XIII

DÉTAILS DE DATE անյայտ, սակայն դատելով ներքին հանգամանքներէն հաւա
առաջին կէսին, և կամ ԺԳ. դին առաջին քառորդին. վասն ս
յիշատակարաններն (տես Թղ. 123բ, 114բն). կը կրեն առաջի
-1377) թուականները, և որոնք համեմատութեամբ երկաթա

LIEU DE COPIE անյայտ, հաւանօրէն Ս. Աստուածածին կամ Ս. Յովհաննէս վ

RÉFÉRENCE t. II, 1924, c. 35-66

GENRE

NOMBRE DE PAGES 494-496

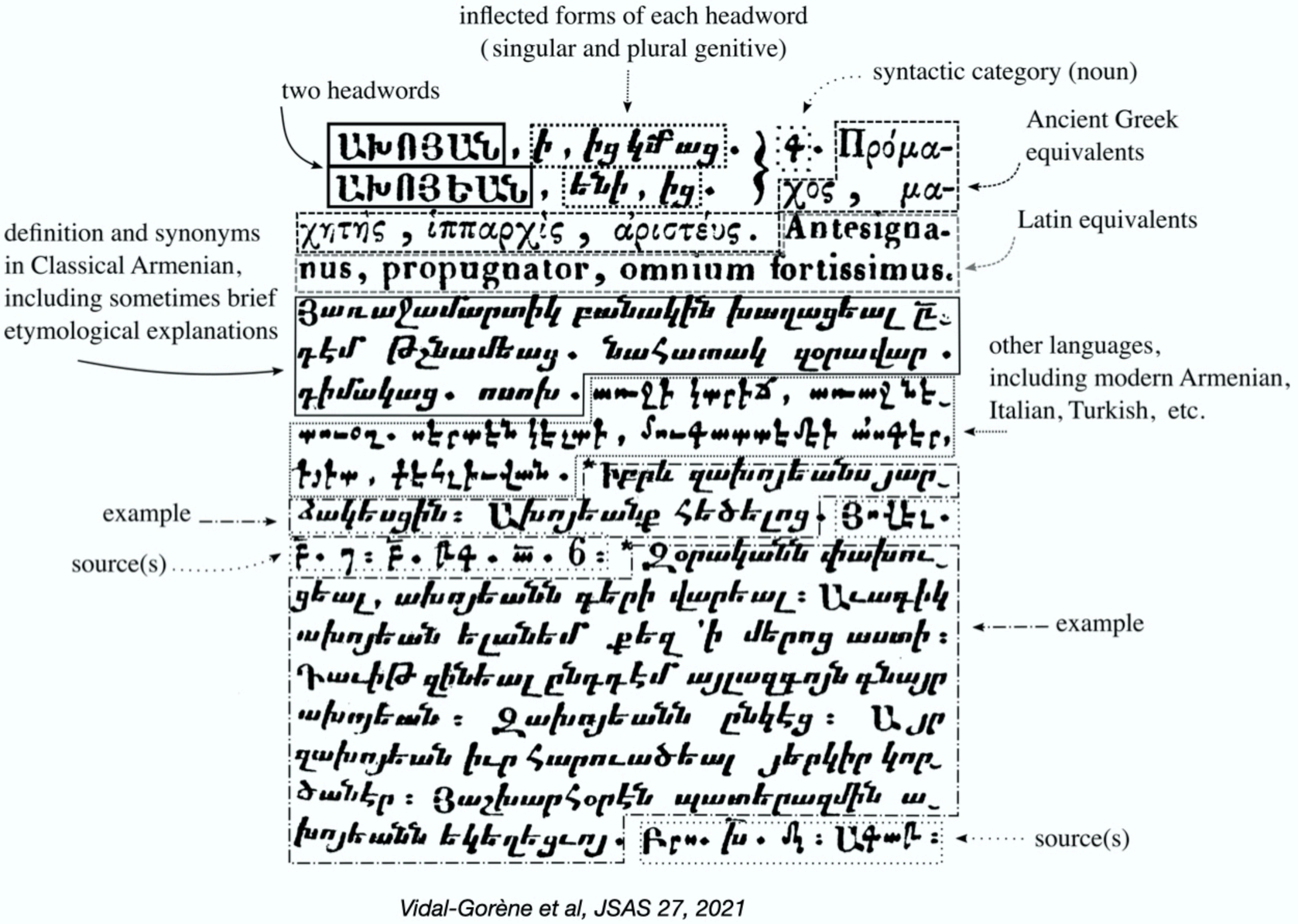
DIMENSIONS 33,4X24,5 (27X21)

MISE EN PAGE երկսիւն, իւրաքանչիւրն 27X9

TYPE D'ÉCRITURE 1

Catalog of Armenian Manuscripts of Venice

5 000 images processed with a model of layout analysis (fine-tuned from newspaper's model) and **OCR content-based segmenter**




*Vidal-Gorène et al, JSAS 27, 2021*

# Ախոյան, ի, ից, աց

Article | NBHL | Étymologie | Mots dérivés

◆ ԱԽՈՅԵԱՆ. πρόμαχος, προμαχητής , ἱππαρχίς, ἀριστεύς
Յառաջամարտիկ բանակին խաղացեալ ընդդէմ թշնա
*կտրիճ, առաջ նետուող. սէրտէն կէչտի, մուգատէմէի*

❖ «Իբրեւ զախոյեանս հարձակեսցին: Ախոյեանք

❖ «Զօրականն փախուցեալ, ախոյեանն գերի վար
զինեալ ընդդէմ այլազգոյն գնայր ախոյեան: Զախո
կործանէր: Յաշխարհօրէն պատերազմին ախոյե
կթ.: Շար.:)

❖ «Ընդդէմ խրոխտ ախոյենին (կամ ախոյանին): Ո
Կործանին վատանուն յախոյանէ անտի: Ի փախու
ախոյանից: Նախամարտիկ եւ առաջին ախոյանի
Եզնիկ.: Պիտ.: Սարկ. քհ.:)

❖ «Մատնեցաւ ի ձեռս ախոյեան թշնամոյն.» (Խոսր

❖ «Առ զյաղթութիւն ի վերայ ախոյանին իւրոյ ... Յա

◆ Նմանութեամբ ասի.

❖ «Զօրեղ ախոյեանս աստուածականս: Ախոյեան
ներհակաց.» (Նար.:)

❖ «[illegible]

Hackathon with the BULAC library and GIS MOMM from January to April 2021 with 15 volunteers

Database of 250 images and annotations will be released in June 2021

Fine-tuning of a model of newspapers with three classes :
text, tables and marginalia. Shown example corresponds to the batch 75

From 16/02/2021 to 16/03/2021 for Armenian Newspapers and Maghrebi Arabic manuscripts :

- 1,081 images have been annotated and proofread
- 1,875 text-regions
- 12,551 lines of text

**Thank you**

CALFA

To know more about the project of Arabic manuscripts : https://www.bulac.fr/espace-recherche/la-bulac-et-la-recherche/

About the platform : https://calfa.fr/vision and process : https://www.aclweb.org/anthology/2020.lrec-1.385.pdf

FONDATION CALOUSTE GULBENKIAN

Fondation des Frères Ghoukassiantz

FÉDÉRATION FRANÇAISE DES TÉLÉCOMS

ՀԻՄՆԱՐԱՐ ՀՀ ԳԱԱ ԳՐԱԴԱՐԱՆ

BULAC
Bibliothèque universitaire des langues et civilisations

Moyen-Orient et Mondes Musulmans
Groupement d'Intérêt Scientifique

cnrs

MEKHITAR