

LE PROJET GRE_gORI

Valorisation des textes écrits dans les langues de l’Orient chrétien (grec, arménien, géorgien, syriaque, arabe, etc.)

Bastien Kindt – Emmanuel Van Elverdinghe

UCLouvain – INCAL – CIOL – Institut orientaliste de Louvain (Louvain-la-Neuve, Belgique)

Contexte

Objectif

Constituer et partager des corpus de textes écrits dans les principales langues de l’Orient chrétien (grec, arménien, géorgien, syriaque, arabe) ainsi que leurs traductions dans les langues modernes.

Challenge

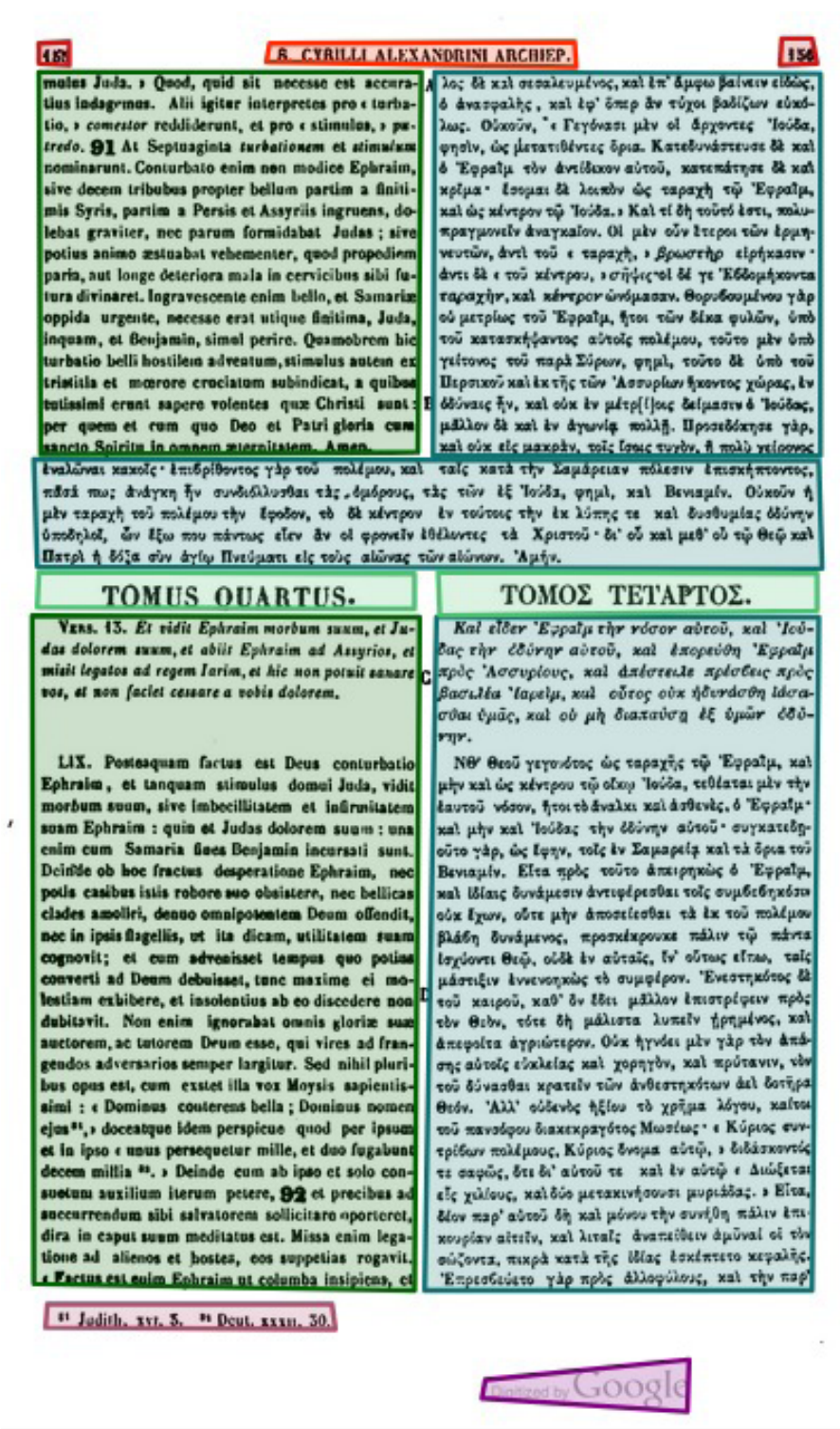
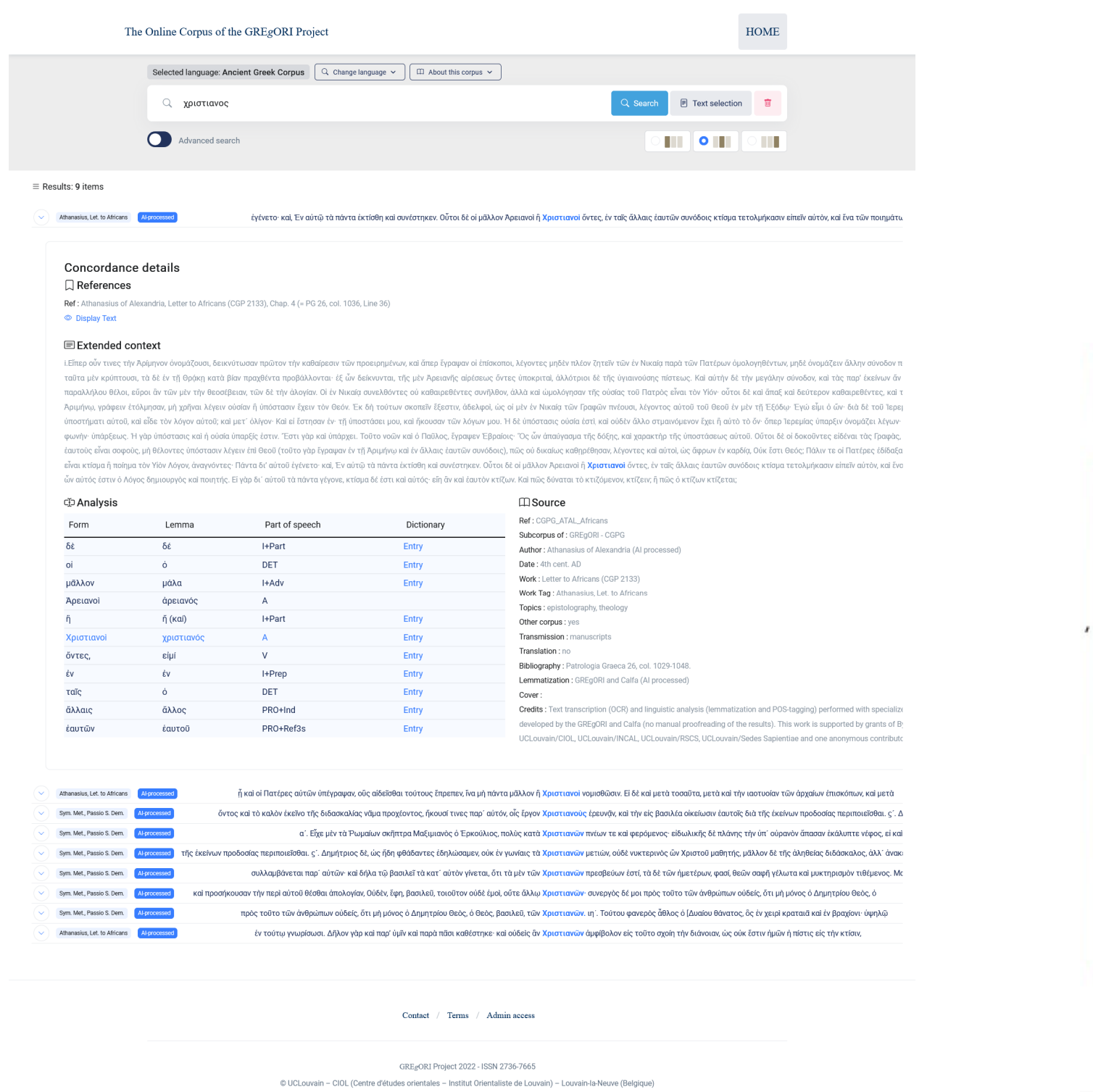
Une aire culturelle et confessionnelle, située géographiquement de la Méditerranée orientale au Caucase, depuis l’Antiquité jusqu’à nos jours. Une abondante production littéraire. Une importante diversité linguistique.

Assurer le traitement et la valorisation de ces textes, malgré :
- différentes familles linguistiques : indo-européenne (grec, arménien), sud-caucasique (géorgien) ou sémitique (syriaque, arabe).
- différentes typologies morphologiques : fusionnelle (grec), agglutinante (géorgien) ou mixte (arménien, syriaque, arabe).
- différents alphabets, orientés de gauche à droite (grec, arménien, géorgien) ou de droite à gauche, avec notation imparfaite des voyelles (syriaque, arabe).
- langues encore relativement peu dotées (rareté des données interoperables et des ressources libres).

Quelques réalisations

Les interfaces du projet GRE_gORI

Accès en ligne, libre et gratuit, aux corpus



Ressources et évaluations

Dictionnaires

GRC – HYE – KAT – SYC – ARA
FRA – DEU – ENG – ITA
Formes, lemmes, catégories (POS-tags)

...

բարոյախաւսեի,բարոյախաւսեմ.V:El1s:Ml1s
բարոյախաւսեին,բարոյախաւսեմ.V:El3p:Ml3p
բարոյախաւսեիք,բարոյախաւսեմ.V:El2s:Ml2s
բարոյախաւսեիք,բարոյախաւսեմ.V:El2p:Ml2p
բարոյախաւսել,բարոյախաւսեմ.V:WAs:WNs:WUs
բարոյախաւսելոյ,բարոյախաւսեմ.V:WÁs:WDs:WGs
բարոյախաւսելով,բարոյախաւսեմ.V:KHs:WHs
բարոյախաւսելովք,բարոյախաւսեմ.V:KHp
բարոյախաւսելոց,բարոյախաւսեմ.V:KÁp:KDp:KF:KGp
բարոյախաւսելում,բարոյախաւսեմ.V:KDs:Kus
...

...

Vérités de terrain

GRC : > 3.000.000 tokens
HYE : 122.000 tokens
KAT : 175.000 tokens
SYC : 223.000 tokens
ARA : 128.000 tokens

Modèles IA (RNN)

Accuracy - all tokens	lemmata	POS	roots
GRC [1]	97.94	98.59	
HYE [2]	93.06	92.38	
KAT [2]	96.98	97.18	
SYC [2]	88.77	88.33	
ARA	86.25	84.14	70.64

[1] Bastien KINDT – Chahan VIDAL-GORÈNE – Saulo DELLE DONNE,
Analyse automatique du grec ancien par réseau de neurones.
Évaluation sur le corpus De Thessalonica Captā. BABELAO 10–11 (2022), p. 537–562.

[2] Chahan VIDAL-GORÈNE – Bastien KINDT, *Lemmatization and POS-tagging process by using joint learning approach. Experimental results on classical Armenian, old Georgian, and Syriac.* LT4HALA 2020, p. 22–27.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	ref	form	ok-lemma	ok-pos	rnn-lemma	rnn-pos	dic-lemma-1	dic-pos-1	dic-lemma-2	dic-pos-2	dic-lemma-3	dic-pos-3
2	58_2_1_9	ել		ել	I+Conj	ել	I+Conj	ել	I+Conj				
3	59_2_1_9	աստ		ասեմ	V	ասեմ	V	ասեմ	V				
4	60_2_1_9	եւս		եւս	N+Ant	եւս	I+Adv	եւս	N+Ant				
5	61_2_1_9	ցկլամ.		ց@կլամ	I+Prep@N+Ant	ցալամ	N+Com	ց@կլամ	I+Prep@N+Ant				
6	62_2_1_9	տէր		տէր	N+Com	տէր	N+Com	տէր	N+Com				
7	63_2_1_9	իմ		էւ	PRO+Per1s	էւ	PRO+Per1s	էւ	PRO+Per1s	իմ	PRO+Pos1s		
8	64_2_1_9	բաղեայ.				բաղեայ	N+Com						
9	65_2_1_9	արի				արի	N+Com	յառեմ	ս	յառեմ	V		
10	66_2_1_9	խնդեմ		խնդեմ	V	խնդեմ	V	խնդեմ	V				
11	67_2_1_9	կերակոյր		կերակոյր	N+Com	կերակոյր	N+Com	կերակոյր	N+Com				
12	68_2_1_9	զի		զի	I+Conj	զի	I+Conj	զի	I+Conj	զի*	PRO+int	զ@ի	I+Prep@I+Prep
13	69_2_1_9	ապրեցոյք.		ապրիմ	V	ապրիմ	V						
14	70_2_1_10	էւ		էւ	I+Conj	էւ	I+Conj	էւ	I+Conj				

DISTAM+ – Paris, 18 mai 2026



Traitements et valorisation

1. Acquisition des textes

- texte manuscrit : HTR
- texte imprimé : OCR
- texte numérisé : encodage, formatage, balisage
- préparation des métadonnées

2. Analyses – lemmatisation et POS-tagging

- approche par dictionnaires et règles
- approche par IA (RNN)
- approche hybride

3. Contrôle des résultats

- environnement de visualisation des résultats de l'analyse hybride
- validation des résultats par un expert
- relecture de concordances de travail et amélioration des données

4. Valorisation des résultats

- mise à jour des ressources (dictionnaires, vérités de terrain, modèles IA, etc.)
- génération d'index
- génération de concordances
- génération de sorties XML-TEI
- constitution de corpus
- mise en ligne de corpus
- mise à disposition des vérités de terrain (GitHub, Zenodo)

Perspectives

- accroître et améliorer les données, les ressources et les modèles (RNN et LLM)
- privilégier l'utilisation de l'approche par LLM pour les traductions
- accroître le volume des données partagées sur les interfaces et les services d'hébergement de données et d'archives

Références

Contact : info-gregori@uclouvain.be
Interfaces : https://www.gregoriproject.com

Développements IA : Calfa



Développements WEB : WebFabrick



Le corpus FLOS (Univ. Ca' Foscari, Venise) – SYC – 109.000 mots.

Un corpus réunissant les textes de florilèges syriaques des citations des Pères grecs. Analyse. Corpus en ligne sur les interfaces de GRE_gORI.

Le corpus eCheikho (Univ. Saint-Joseph, CEDRAC, Beyrouth) – ARA – 128.000 mots.

Création d'un corpus d'entraînement pour le développement d'un modèle IA (RNN) dédié à l'analyse des textes arabes chrétiens. Vérité de terrain en ligne sur les interfaces de GRE_gORI.

Le projet Digital Armenian Colophons (UCLouvain/FRS-FNRS) – HYE – 1.725.000 mots.

Un corpus des colophons de manuscrits arméniens. Acquisition et analyse (RNN). Traitement des entités nommées. Base PostgreSQL pour l'ensemble des métadonnées.