Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm 2022) @LREC2022, pages 20–27
Marseille, 20 June 2022

Analyse Automatique de l'Ancien Arménien. Évaluation d'une méthode hybride « dictionnaire » et « réseau de neurones » sur un Extrait de l'*Adversus Haereses* d'Irénée de Lyon

Gabriel Kepeklian, Bastien Kindt
Centre d'études orientales – Institut orientaliste de Louvain (CIOL),
Institut des civilisations, arts et lettres (INCAL),
UCLouvain, Louvain-la-Neuve, Belgium
{gabriel.kepeklian,bastien.kindt}@uclouvain.be

Abstract

The aim of this paper is to evaluate a lexical analysis (mainly lemmatization and POS-tagging) of a sample of the ancient-Armenian version of the *Adversus Haereses* by Irenaeus of Lyons (2nd c.) by using hybrid approach based on digital dictionaries on the one hand, and on Recurrent Neural Network (RNN) on the other hand. The quality of the results is checked by comparing data obtained by implementing these two methods with data manually checked. In the present case, 98,37% of the results are correct by using the first (lexical) approach, and 74,64% by using the second (RNN). But, in fact, both methods present advantages and disadvantages and argue for the hybrid method. The linguistic resources implemented here are jointly developed and tested by GREgORI and Calfa.

Mots-clés : ancien arménien, lemmatisation, étiquetage morphosyntaxique (POS-tagging), réseau de neurones (RNN)

1. Introduction

1.1 Irénée de Lyon et l'*Adversus Haereses*

Irénée (mort vers 202 ap. J.-C.) est le deuxième évêque de Lyon (Lugdunum), capitale des trois Gaules, territoires alors soumis à l'Empire romain. Père de l'Église, il est aussi considéré comme le premier théologien. Natif de Smyrne en Asie Mineure, sa langue et sa culture sont grecques. Son œuvre principale, écrite en grec, est une *Présentation et réfutation de la gnose au faut nom* (*Ἔλεγχος ἀνατροπὴ τῆς ψευδωνύμου γνώσεως*), en cinq livres. L'auteur y réfute les doctrines gnostiques venues d'Asie Mineure, puis y développe une riche pensée théologique (Rousseau, 1984). Ce texte est perdu, mais deux traductions sont parvenues jusqu'à nous. La première est une traduction latine (IV^e^-V^e^ s.) transmise sous le titre (réducteur) d'*Adversus Haereses* (« Contre les Hérésies » ; désormais *AH*). La seconde est une traduction arménienne (VII^e^ s.) intitulée *Յանդիմանութեան եւ եղծման ստաւուն գիտութեանն Yandimanut'ean ew ełcman stanun gitut'eann* (« Présentation et réfutation de la gnose au faux nom », traduction du titre grec) connue par un unique manuscrit du XIII^e^ s., conservé au Maténadaran à Erevan sous la cote M3710. De nombreux fragments grecs et arméniens, et quelques autres latins et syriaques complètent en outre ces deux traductions. Le cinquième livre de l'*AH*, dans sa version arménienne, vient de faire l'objet d'une nouvelle édition par (Kepeklian, 2021)[1]. L'analyse lexicale de ce texte est en cours dans le cadre du projet GREgORI[2]. Le présent article porte sur un extrait de ce livre V qui s'étend de la préface au chapitre II, 3. Le tableau 1 indique le nombre de mots-occurrences et le nombre de formes différente dans l'ensemble du livre V et, pour l'extrait, qui est déjà analysé, le nombre de lemmes.

	Mots-occurrences (tokens)	Formes de mots (unique token)	Lemmes
Livre V	25.544	6.069	(en cours d'analyse)
Préface - Ch. II, 3	1.530	756	444

Tableau 1 : Nombre de mots-occurrences, de formes de mots et de lemmes dans l'*AH*, V et dans l'extrait (Préface - Ch. II, 3)

1.2 L'analyse de l'extrait de l'*AH* et les ressources linguistiques du projet GREgORI pour l'ancien arménien

À terme, l'analyse du livre V de l'*AH* fournira aux chercheurs un corpus entièrement étiqueté de ce texte. Ce corpus sera accessible en ligne via les interfaces du projet GREgORI[3]. Cette analyse comprend la lemmatisation de tous les mots du corpus ainsi que leur étiquetage morphosyntaxique (POS) et flexionnel. Deux méthodologies sont adoptées pour réaliser la lemmatisation et les étiquetages. 1) Une première approche compare le vocabulaire du texte aux lexiques de référence des ressources linguistiques du projet GREgORI. 2) Une seconde approche utilise un réseau de neurones préparé par Calfa[4]. Cette démarche hybride alliant une approche dite « par dictionnaires » et une approche ayant recours à l'« intelligence artificielle » est désormais privilégiée pour l'analyse des textes arméniens, géorgiens, grecs et syriaques traités dans le cadre du projet GREgORI (Vidal-Gorène and Kindt, 2020 ; Vidal-Gorène and Kindt, 2022 ; Kindt, Vidal-Gorène et Delle Donne, 2022). Pour le moment, seule la première des deux approches, celle par « par dictionnaires », fournit les analyses flexionnelles.

[1] Cette nouvelle édition sera publiée dans le *Corpus Scriptorum Christianorum Orientalium* édité par Peeters Publishers (Leuven, Belgique).

[2] Le projet GREgORI est mené à l'Institut orientaliste de l'UCLouvain, sous la direction du professeur Bernard Coulie ; https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html

[3] https://www.gregoriproject.com

[4] https://calfa.fr

1.2.1 Les lexiques de référence

Comme l'illustre le tableau 2, les ressources lexicales du projet GREgORI pour l'arménien sont réunies dans des lexiques de référence totalisant 1.199.123 formes de mots, regroupées sous 30.311 lemmes.

Formes simples	315.952
Formes composées	883.171
Nombre total de formes	1.199.123
Lemmes	30.311

Tableau 2 : Nombre de lemmes, de formes simples et de formes composées enregistrées dans les ressources linguistiques du projet GREgORI

Ces ressources distinguent les formes dites « simples » (1) et les formes dites « composées » (2).

1) մարդ *mard* « homme », lemme մարդ, catégorie morphosyntaxique N+Com (nom commun), analyses flexionnelles :As:Ns:Us (accusatif singulier, nominatif singulier et locatif singulier) (les étiquettes morphosyntaxiques et flexionnelles sont énumérées dans les tableaux les annexes 3 et 4, cfr 7.3 et 7.4).

2) զմարդն *zmardn* « l'homme », segmenté lors de l'analyse en զ-մարդ-ն, I+Prep (préfixe prépositionnel), N+Com (nom commun) et PRO+Dem (suffixe déterminatif), :As (accusatif singulier)[5].

La distinction établie entre formes simples et formes composées, et donc la discrimination des préfixes prépositionnels et des suffixes déterminatifs, permet d'inclure ces éléments lexicaux dans les analyses. Ces éléments peuvent donc servir d'arguments dans les requêtes formulées par les chercheurs explorant le corpus. Les principes de formulation des intitulés de lemme et les étiquettes morphosyntaxiques et flexionnelles sont décrits dans (Coulie, Kindt, Kepeklian et Van Elverdinghe, 2022).

Les formes enregistrées dans les ressources sont soit des formes effectivement attestées dans le corpus complet des textes déjà traités dans le cadre du projet – ensemble textuel décrit dans (Vidal-Gorène, Ch. and Kindt, 2020 ; le tableau 3 en indique les effectifs) –, soit des formes générées automatiquement avec une supervision par un expert humain, comme expliqué dans (Coulie, Kindt, Kepeklian, et Van Elverdinghe, 2022).

Mots-occurrences (tokens)	73.211
Formes simples	61.291
Formes composées	11.920
Formes de mots (unique tokens)	17.554
Formes simples	11.851
Formes composées	5.703
Lemmes	5.649

Tableau 3 : Effectifs des formes effectivement attestées dans les textes déjà traités

1.2.2 Le réseau de neurones

L'approche par réseau de neurones est basée sur un apprentissage mis en œuvre sur le corpus des textes déjà traités. Elle a déjà été testée et évaluée, en arménien comme dans d'autres langues de l'Orient chrétien (Vidal-Gorène and Kindt, 2020 ; Vidal-Gorène and Kindt, 2022). Le tableau 4 rappelle les résultats obtenus à cette occasion sur un corpus de test. Sur l'ensemble des mots du corpus de test, l'*accuracy* atteint 0.9044 pour la lemmatisation et 0.9238 pour l'étiquetage morpho-syntaxique (résultats de mai 2020). Pour rappel, cette approche ne fournit pas encore les informations flexionnelles. Deux constats ont été établis lors de cette évaluation : 1) les résultats observés sont meilleurs sur les formes ambiguës que sur les formes inconnues ; 2) les résultats sont meilleurs pour la catégorisation morphosyntaxique que pour la lemmatisation.

	Toutes les formes (tokens)	Formes ambiguës (tokens)	Formes inconnues (tokens)
Lemmatisation			
accuracy	0.9044	0.8620	0.6864
precision	0.6630	0.4411	0.5074
recall	0.6711	0.5211	0.5118
f1-score	0.6670	0.4778	0.5096
Étiquetage morphosyntaxique			
accuracy	0.9238	0.9145	0.7441
precision	0.6513	0.6306	0.2920
recall	0.6264	0.6501	0.3124
f1-score	0.6386	0.6402	0.3019

Tableau 4 : Résultats de la lemmatisation et de l'étiquetage morphosyntaxique par réseau de neurones

1.2.3 Objectif de cette contribution

Le but de cet article est d'évaluer une nouvelle fois les résultats acquis par les deux approches, celle basée sur l'utilisation des ressources du projet GREgORI (désormais GREgORI) et celle basée sur un réseau de neurones (désormais RNN, pour l'anglais *Recurrent Neural Network*).

Il faut noter que l'analyse de GREgORI ne tient pas compte du contexte et est entièrement dépendante du contenu des ressources linguistiques mises en œuvre. Cette approche fournit une ou plusieurs analyses possibles pour les mots du texte connus des ressources, mais aucun résultat pour les mots inconnus des ressources. *A contrario*, l'analyse par RNN tient compte du contexte d'apparition des mots dans le texte et propose une analyse pour tous les mots, qu'ils soient univoques, équivoques, ou inconnus du corpus d'apprentissage. Dans l'expression ի ձեռն հոգւոյն աստուածոյ *i jeřn hogwoyn astuacoy* « grâce à l'esprit de Dieu », GREgORI fournit pour la forme ի les quatre lemmes possibles hors contexte, à savoir ի *i* (la préposition), իւի *ini* (le nom de la lettre), 20 (pour le déterminant cardinal) et 20th (pour le déterminant numérique ordinal). Dans ce cas, le RNN prédit à juste titre une seule analyse : ի *i* (la préposition). En utilisant les ressources de GREgORI, l'occurrence մարդս *mards* « homme » ou « hommes » – attestée dans l'expression եւ զինչ մարդս բարեգործի *ew zinč' mards baregorci* « et en quoi l'homme est objet de bienfaits » –, reçoit deux analyses : մարդ.N+Com:Ap:Up (une forme « simple » à l'accusatif ou au locatif pluriel) et մարդ@u. N+Com@PRO+Dem:As:Ns:Us@Ø (forme « composée »

[5] Tous les exemples arméniens cités sont tirés de l'extrait de l'*AH.*

à l'accusatif singulier, au nominatif singulier ou au locatif singulier munie du suffixe déterminatif -u). Ici, le RNN prédit à tort la forme simple.

Quand elles sont univoques, les analyses de GREgORI sont très fiables. En revanche, une révision par un expert humain reste nécessaire pour achever l'analyse des formes inconnues et ambiguës. Les analyses produites par RNN sont quant à elles des prédictions. Pour fournir un corpus parfaitement étiqueté, une révision par un expert humain est, une fois encore, indispensable. Mais l'arménien ancien reste une langue peu-dotée (Vidal-Gorène and Decours-Perez, 2020) et il semble utile de conserver les deux types d'analyse. La complémentarité des approches peut dès lors s'appréhender en considérant les dimensions de leurs zones d'ombre conjointes. Dans l'extrait de l'*AH*, GREgORI ne fournit aucune analyse pour vingt mots (soit 1,3%). Pour dix d'entre eux, RNN propose correctement le lemme et la catégorie morphosyntaxique (annexe 1, cfr 7.1). Pour six autres (soit moins de 0,4%), RNN ne propose ni le bon lemme ni la bonne catégorie morphosyntaxique (annexe 2, cfr 7.2).

Il est possible d'expliquer pourquoi GREgORI ne fournit aucune analyse pour les dix formes consignées dans l'annexe 1, cfr 7.1 :

- les trois formes գնացելում, եղելում et յաղթեցելում sont des participes post-classiques, au datif ou au locatif ;

- la présence du déterminatif -ն en finale des deux formes երեւէրն, կամէրն se justifie par le fait que ces verbes constituent les deuxièmes termes d'une proposition relative.

Ces différentes formes et différents usages ne sont pas systématiquement décrits dans les ressources du projet. Quant au verbe կացուցանեմ, il n'est tout simplement pas encore enregistré dans les lexiques de référence.

2. Évaluation

Disposant de deux approches foncièrement différentes, il est particulièrement intéressant de les confronter. L'évaluation reposera sur la comparaison des résultats de GREgORI et de RNN à ceux d'une révision manuelle (désormais Révision), car l'échantillon considéré est déjà analysé et a fait l'objet d'un premier contrôle. Dans les lignes qui suivent, nous abordons la combinatoire des situations d'accord et de désaccord entre les différentes approches et nous les illustrons d'exemples.

2.1 Accord entre GREgORI et Révision

Accord sur	Nombre	%
le lemme et la catégorie	1.505	98,37%

Tableau 5 : Accord GREgORI *vs* Révision

Dans la très grande majorité des cas, parmi les analyses fournies par GREgORI (une seule ou plusieurs) se trouve l'analyse correcte, que ce soit pour des formes simples ou composées (cfr 1.2.1).

1) այլ հաստատուն **իրաւք** ճշմարտութեան լինէր արդեւք (*AH* V 1.1), *ayl hastatun irawk' čšmartut'ean linēr ardewk'* « mais assurées par des faits véridiques »

– իրաւք,իր.N+Com:Hp –

la forme իրաւք *irawk'* « faits » a pour lemme իր *ir*, nom commun à l'instrumental pluriel.

2) եթե ոչ **վարդապետն** մեր (…) մարդ եղանիւր (*AH* V 1.1), *et'e oč' vardapetn mer (…) mard eɫaniwr* « si notre maître (…) ne s'était fait homme »

– վարդապետն,.N+Com:As:Ns@ն,.PRO+Dem –

la forme վարդապետն *vardapetn* « maître » est composée de deux éléments dont les lemmes sont վարդապետ *vardapet* et -ն *-n*. Le premier est un nom commun au nominatif singulier, le second un déterminatif.

3) ո՞ այլ ոք գիտաց **զմիտս** Աստուածոյ (*AH* V 1.1) *o? ayl ok' gitac' zmits Astuacoy* « qui d'autre a connu la pensée de Dieu »

– զ,.I+Prep@միտն,.N+Com:Ap –

la forme զմիտս *zmits* « pensée » est composée de deux éléments dont les lemmes sont զ- *z-* et միտն *mit* qui sont respectivement une préposition et un nom commun à l'accusatif pluriel.

2.2 Désaccord entre GREgORI et Révision

Désaccord, GREgORI n'a	Nombre	%
aucune analyse satisfaisante (4)	25	1,63%
ou pas d'analyse (5)	20	1,31%

Tableau 6 : Désaccord GREgORI *vs* Révision

Lorsque GREgORI fournit au moins une analyse, aucune n'est correcte pour vingt-cinq mots. Enfin, GREgORI ne propose aucune analyse pour vingt mots du corpus (on a bien 98,37+1,63 = 100%). Ces deux ensembles de vingt-cinq et vingt mots n'ont, par définition, aucun mot en commun.

4) այլ **ամայի** անապատ եղելոյ (*AH* V 2.1) *ayl amayi anapat eɫeloy* « mais devenu privé »

– ամայի,.A:As:Ns –

la forme ամայի *amayi* « privé » est l'adjectif ամայի au nominatif singulier. Dans les ressources de GREgORI, cette forme n'est enregistrée que sous le verbe ամամ.

5) **բարւոքն** պահեցեալ յեկեղեցւոյ (*AH* V praef.) *barwok'n pahec'eal yekeɫec'woy* « la [foi] bien gardée dans l'église »

– բարւոք,.A:As:Ns@ն,.PRO+Dem –

la forme բարւոքն *barwok'n* « bien » correspond à l'adjectif բարւոք *barwok'* au nominatif singulier suffixé du déterminatif -ն *-n*, lemme absent des ressources de GREgORI.

2.3 Accord entre RNN et Révision

Accord sur	Nombre	%
le lemme (6)	1201	78,50%
la catégorie (7)	1308	85,49%
le lemme et la catégorie (8)	1142	74,64%

Tableau 7 : Accord RNN *vs* Révision

6) **ամենեցուն** որք պատահիցեն գրոյս այսմիկ (*AH* V praef.) *amenec'un ork' patahic'en groys aysmik* « à tous ceux qui rencontreront ce livre »

– ամենեցուն,ամենեքեան.PRO+Ind:Âp:Dp:Gp –

la forme ամենեցուն *amenec'un* « tous » est le pronom indéfini ամենեքեան *amenek'ean* au datif pluriel. RNN a bien prédit le lemme mais le caractérise comme nom commun.

7) nʔ այլ ոք խորհրդակից եղեւ **նորա** (*AH* V 1.1) *oʔ ayl ok' xorhrdakic' eɫew nora* « qui d'autre a été son conseiller ? »

– նորա, նա (նա).PRO+Dem:Gs –

la forme նորա est le génitif singulier du pronom démonstratif նա dont le lemme est նա (նա) afin d'éviter l'homographie avec la conjonction նա (եւ). La forme est fréquente. Pourtant, RNN propose un lemme նա (սա), sans doute présent, erronément, dans le corpus d'apprentissage.

8) ոչ **բռնադատելով** առնուլ զորս կամէրն (*AH* V 1.1) *oč' bṙnadatelov aṙnul zors kamērn* « ne forçant pas à prendre celles qu'il voulait »

– բռնադատելով,բռնադատեմ.V:KHs:WHs –

RNN propose pour la forme բռնադատելով *bṙnadatelov* « forcer », du lemme verbal բռնադատեմ *bṙnadatem.* Il s'agit bien de l'infinitif du verbe à l'instrumental singulier.

2.4 Désaccord entre RNN et Révision

Désaccord sur	Nombre	%
la catégorie, mais accord sur le lemme (9)	59	3,86%
le lemme, mais accord sur la catégorie (10)	166	10,85%
le lemme et la catégorie (11, 12)	163	10,65%

Tableau 8 : Désaccord RNN *vs* Révision

9) **որպէս** երանելի առաքեալն ասէ (*AH* V 2.3) *orpēs eraneli aṙak'ealn asē* « comme le bienheureux apôtre dit »

– որպէս,.I+Conj –

La conjonction որպէս *orpēs* est erronément caractérisé comme un adverbe par RNN, analyse sans doute présente, erronément, dans le corpus d'apprentissage.

10) եւ յայլ եւս ի մարդն **զոյացութենէ** (*AH* V 2.2) *ew yayl ews i mardn goyac'ut'enē* « et du reste de la substance de l'homme »

– զոյացութենէ,զոյացութիւն.N+Com:Âs –

la forme զոյացութենէ *goyac'ut'enē* « substance » doit être comprise comme l'ablatif singulier du nom commun զոյացութիւն *goyac'ut'iwn*. RNN propose un nom commun, mais lui attribue un lemme qui n'existe pas : զոյանութիւն.

11) ի **լաւէն** զառ նա (*AH* V 1.1) *i lawēn zaṙ na* « à partir du bon auprès de lui »

– լաւէ,լաւ,A:Âs@ն,.PRO+Dem –

la forme լաւէն *lawēn* est l'ablatif singulier de l'adjectif լաւ accompagnée du suffixe déterminatif -ն. RNN suggère une analyse possible, à savoir une forme conjuguée du verbe լսեմ. Mais cette prédiction ne convient pas *in textu.*

12) որ (…) ի **հացէ** որ մարմին նորա աճէր (*AH* V 2.2) *or (…) i hac'ē or marmin nora ačēr* « qui (…) s'accroissait par le pain qui est son corps »

– հացէ,հաց.N+Com:Âs –

la forme հացէ est l'ablatif singulier du nom commun հաց *hac'* « pain ». RNN prédit un lemme verbal հաց. Le RNN prédit trente-six lemmes verbaux impropres car ne se terminant pas par -ամ, -եմ, -իմ ou -ում.

2.5 Accord entre GREgORI et RNN

Après avoir aborder chaque outil isolément, nous prenons ici en compte leur accord sur une même analyse. Dans 73,86% des cas la prédiction du RNN correspond à une des analyses possibles proposées par GREgORI. Cet accord peut être correct ou fautif.

Accord sur	Nombre	%
le lemme et la catégorie qui sont corrects (13)	1.130	73,86%

Tableau 9 : Accord GREgORI, RNN *vs* Révision

13) զի եւ զենթադրութիւնս զայս **գիտասցես** (*AH* V praef.) *zi ew zent'adrut'iwns zays gitasc'es* « afin que tu connaisses aussi cet argument »

– գիտասցես,գիտեմ.V:ESJ2s –

les deux outils classent la forme գիտասցես sous le lemme verbal գիտեմ.

Dans l'exemple qui suit, les deux outils s'accordent cependant sur une analyse erronée. Cela peut s'expliquer par le fait que RNN a été entraîné sur les données de GREgORI dans lesquelles cette analyse fautive est présente.

14) Եւ եթե ոչ **ապրեսցի** սա (*AH* V 2.2) *Ew et'e oč' apresc'i sa* « et si elle n'est pas sauvée »

– ապրեսցի,ապրիմ.V:MSJ3s –

Les deux outils s'accordent bien sur la nature verbale du mot, mais propose un lemme actif ապրեմ au lieu de ապրիմ (« se sauver »).

3. Conclusions et perspectives

Les données de GREgORI et de la Révision s'accordent dans 98,37% des cas (cfr 2.1). Cela plaide en faveur de l'analyse produite par GREgORI. Les données de RNN et de la Révision (cfr 2.3) s'accordent dans 74,64% des cas. L'accord entre GREgORI et RNN concerne 73,86% des cas (cfr 2.5). Ces deux derniers résultats sont donc inférieurs au premier. En revanche, quand, à vingt reprises, GREgORI ne fournit aucune analyse, RNN prédit dix analyses correctes, tant au niveau du lemme que de la catégorie morphosyntaxique (cfr 1.2.3). Ces résultats sont illustrés dans la figure 1 (voir aussi l'annexe 1, cfr 7.1).

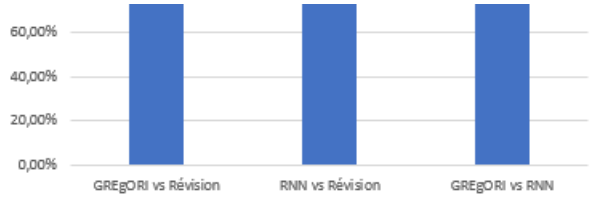

Figure 1 : Accords sur les lemmes
et les catégories morphosyntaxiques

Lorsqu'elles ne s'accordent pas, ces deux approches demeurent complémentaires. Les cas où elles n'ont ni l'une ni l'autre la bonne analyse restent minoritaires : GREgORI *vs* Révision, 1,63% (cfr 2.2), RNN *vs* Révision, 10,65% (cfr 2.4). Ces données sont illustrées dans la figure 2. Par ailleurs, aucun outil n'invalide l'autre et leur utilisation conjointe permet même au réviseur humain de travailler efficacement. Quand, à vingt reprises, GREgORI ne fournit aucune analyse, RNN se trompe six fois sur le lemme et sur la catégorie morphosyntaxique (voir aussi l'annexe 2, cfr 7.2). Les autres résultats sont corrects.

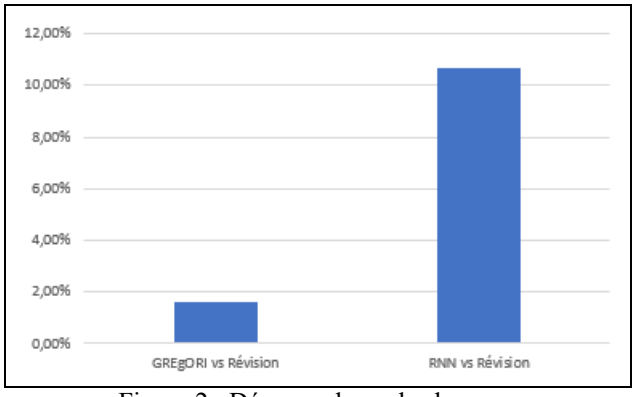

Figure 2 : Désaccords sur les lemmes
et sur les catégories morphosyntaxiques

Les corrections apportées lors de la révision manuelle enrichissent les données. Pour les ressources de GREgORI, ce sont des ajouts de formes nouvelles, simples ou composées, ou de lemmes inédits, incluant les informations morphosynatxiques et flexionnelles correspondant aux formes concernées. Il y a aussi des corrections. Pour RNN, les données lemmatisées de chaque nouveau texte traité rejoignent, après révision, le corpus d'apprentissage utilisé pour construire, tester et évaluer le réseau de neurones, avant son utilisation sur de nouveaux textes (cfr 1.2.2).

Plus ces outils seront utilisés, meilleurs ils seront. Le projet GREgORI est basé sur les itérations successives de ses outils et sur leur hybridation. Plusieurs corpus sont actuellement en cours de traitement ou de révision. L'examen des analyses produites ou prédites à l'occasion du traitement de ces textes permettra d'objectiver l'évolution progressive des performances de ces deux approches et de les comparer aux résultats acquis précédemment. À court ou moyen termes, l'accroissement des données déjà analysées permettra de paramétrer le RNN pour qu'il prédise aussi les analyses flexionnelles.

Après les inévitables phases de développement, d'implémentation et de test (comme décrit dans Vidal-Gorène and Kindt, 2020 ; Vidal-Gorène and Kindt, 2022 ; Kindt, Vidal-Gorène et Delle Donne, 2022), l'approche hybride combinant les analyses « par dictionnaire » et par « réseau de neurones » entre dans une phase de réelle production, en arménien, mais aussi dans les autres langues de l'Orient chrétien (géorgien, syriaque,grec, etc.). Outre le livre V de l'*AH* d'Irénée, en cours de traitement sous la responsabilité de Gabriel Kepeklian (cfr 1.1 et note 1), les textes arméniens de deux volumes du CSCO ont ou vont bientôt rejoindre les données lemmatisées de GREgORI. Bernard Coulie a analysé le *Commentaire à la Genèse* attribué à Step'anos de Siwnik' (CSCO 695, Scrip. Arm. 32) publié par M.E. Stone, ainsi que la version arménienne des *Lettres* d'Évagre le Pontique (CSCO 704, Scrip. Arm. 33) publiée par R. Darling Young et H. Karapetyan. Par ailleurs, l'analyse de tous les volumes arméniens du CSCO est en cours, en collaboration avec Peeters Publishers. Emmanuel Van Elverdinghe assure l'analyse des trois versions arméniennes déjà éditées le l'*Apocalypse* de Jean, (Murat, 1905 ; Conybeare, 1907 ; Zōhrapean, 1805), ainsi que celle des textes des Colophons des manuscrits arméniens (Van Elverdinghe, 2018 ; Van Elverdinghe, 2022 ; Van Elverdinghe et Kindt, 2022).

4. Remerciements

Les auteurs tiennent à exprimer leur gratitude envers le Professeur Bernard Coulie (UCLouvain), Chahan Vidal-Gorène (Calfa), et Emmanuel Van Elverdinghe (UCLouvain).

5. Bibliographical References

Conybeare, F.C. (1907). The Armenian Version of Revelation and Cyril of Alexandria's Scholia on the Incarnation and Epistle on Easter. *Text and Translation Society*, 5. London : p. 1-32.

Coulie, B., Kindt, B., Kepeklian, G. & Van Elverdinghe, E. (2022). Étiquettes morphosyntaxiques et flexionnelles pour le traitement automatique de l'arménien ancien. *Le Muséon*, 135 (1-2): p. 209-241.

Kepeklian, G. (2021). La version arménienne du Livre V de l'*Adversus haereses* d'Irénée de Lyon : histoire du texte, édition critique, traduction et notes (Thèse de doctorat), UCLouvain, Louvain-la-Neuve.

Kindt, B., Vidal-Gorène, Ch. & Delle Donne, S. (2022). Analyse automatique du grec ancien par réseau de neurones. Évaluation sur le corpus *De Thessalonica Capta*. *BABELAO*, 10-11: p. 537-562.

Murat, Fr. (1905-1911). Յայտնութեանն Յովհաննու հին հայ թարգմանութիւն *(Yaytnut'eann Yovhannu hin hay t'argmanut'iwn)* / *Die Offenbarung Johannis in einer alten armenischen Übersetzung*). Jerusalem: p. 3-76.

Rousseau, A. (1984), Irénée de Lyon, Contre les hérésies. Dénonciation et réfutation de la gnose au nom menteur (Sagesses Chrétiennes). Paris: Les éditions du Cerf.

Van Elverdinghe, E. & Kindt, B. (2022). Describing Language Variation in the Colophons of Armenian Manuscripts. LREC 2022. Submitted.

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

Van Elverdinghe, E. (2022). Modèles et copies. Étude d'une formule des colophons de manuscrits arméniens (VIII^{e}-XIX^{e} siècles). Louvain: Peeters.

Vidal-Gorène, Ch. & 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*, p. 3145-3152). Paris: The European Language Resources Association (ELRA).

Vidal-Gorène, Ch. & 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 & M. Passarotti (Eds.), *1st Workshop on Language Technologies for Historical and Ancient Languages, (LT4HALA 2020): Proceedings*, p. 22-27. Paris: European Language Resources Association (ELRA).

Vidal-Gorène, Ch. & 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.

Zōhrapean, Y. (1805). Ա(ստուա)ծաշունչ մատեան հին եւ նոր կտակարանաց (*Astuacašunč' matean hin ew nor ktakaranac' / God-Breathed Scriptures of the Old and New Testaments*), Venice : p. 825-836.

6. Language Resource References

Calfa. (depuis 2014). Calfa, https://calfa.fr.

GREgORI Project. (since 1990). GREgORI – Software, linguistic data and tagged corpora for ancient GREek and ORIental languages, https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html, ISSN 2736-7657.

7. Annexes

7.1 Annexe 1 : Liste des formes inconnues de GREgORI mais analysées correctement par RNN

Forme (token)	Lemme	Catégorie morphosyntaxique
գնացելում	գնամ	V
եղելում	եղանիմ	V
զաստուածոյսն	զ@աստուած	I+Prep@N+Com@PRO+Dem
կամէրն	կամիմ@ն	V@PRO+Dem
երեւէրն	երեւիմ@ն	V@PRO+Dem
երեւէրն	երեւիմ@ն	V@PRO+Dem
յաղթեցելում	յաղթեմ	V
կացուսցէ	կացուցանեմ	V
զվերստին	զ@վերստին	I+Prep@I+Adv
զանդրէն	զ@անդրէն	I+Prep@I+Adv

Cfr explication en 1.2.3.

7.2 Annexe 2 : Liste des formes inconnues de GREgORI et analysées erronément par RNN

Forme (token)	RNN		Révision	
	Lemme	Catégorie morphosyntaxique	Lemme	Catégorie morphosyntaxique
կատարելոյն	կատարելի	A	կատարելի@ն	A@PRO+Dem
այժմս	այժմ	I+Adv	այժմ@u	I+Adv@PRO+Dem
բարւոքն	բարւոք@ն	NUM+Ord	բարւոք@ն	A@PRO+Dem
բարձրելոյն	բարձրեալ	A	բարձրեալ@ն	N+Com@PRO+Dem
շարունակէն	շարունական	A	շարունակեմ@ն	V@PRO+Dem

Cfr explication en 2.4.

7.3 Annexe 3 : Liste des étiquettes morphosyntaxiques (POS) (tiré de Coulie, Kindt, Kepeklian et Van Elverdinghe, 2022)

Étiquette	Description	Étiquette	Description
A	Adjectif	NUM+Car	Déterminant numérique cardinal (mot)
I+Adv	Mot invariable – Adverbe	NUM+Ord	Déterminant numérique ordinal (mot)
I+AdvPr	Mot invariable – Adverbe prépositionnel	NUMA+Car	Déterminant numérique cardinal (lettre)
I+Conj	Mot invariable – Conjonction	NUMA+Ord	Déterminant numérique ordinal (lettre)
I+Intj	Mot invariable – Interjection	PRO+Dem	Pronom démonstratif
I+Neg	Mot invariable – Négation	PRO+Ind	Pronom indéfini
I+Part	Mot invariable – Particule	PRO+Int	Pronom interrogatif
I+Prep	Mot invariable – Préposition	PRO+Per[1,2][s,p]	Pronom personnel
N+Ant	Nom propre anthroponymique	PRO+Pos[1,2][s,p]	Pronom possessif
N+Com	Nom commun	PRO+Rec	Pronom réciproque
N+Let	Nom d'une lettre	PRO+Ref	Pronom réfléchi
N+Pat	Nom propre patronymique	PRO+Rel	Pronom relatif
N+Prop	Nom propre	V	Verbe
N+Top	Nom propre toponymique		

7.4 **Annexe 4 : Liste des étiquettes flexionnelles** (tiré de Coulie, Kindt, Kepeklian et Van Elverdinghe, 2022)

Type d'étiquette	Étiquette	Description	Type d'étiquette	Étiquette	Description
Cas	N	Nominatif	Mode	Î	Indicatif
	A	Accusatif		K	Participe
	G	Génitif		S	Subjonctif
	D	Datif		Y	Impératif
	U	Locatif		W	Infinitif
	Â	Ablatif	Temps	P	Présent
	H	Instrumental		I	Imparfait
Nombre	s	Singulier		J	Aoriste
	p	Pluriel	Personne	1	Première personne
Voix	E	Actif		2	Deuxième personne
	B	Passif		3	Troisième personne
	M	Moyen-passif			