ÉTIQUETTES MORPHOSYNTAXIQUES ET FLEXIONNELLES POUR LE TRAITEMENT AUTOMATIQUE DE L'ARMÉNIEN ANCIEN

1. *Introduction*

1.1. *Le projet GREgORI*

Les travaux du projet GRE*g*ORI menés à l'Institut orientaliste de l'Université catholique de Louvain (UCLouvain, Belgique) portent sur des textes écrits en grec ancien et dans les langues de l'Orient chrétien, en particulier l'arabe, le syriaque, le géorgien ancien et l'arménien ancien[1]. L'objectif du projet est de mettre à la disposition des chercheurs des corpus de textes enrichis d'un étiquetage grammatical indiquant, pour chaque forme du corpus, son lemme, c'est-à-dire une entrée de dictionnaire – յուսով (yusov) ayant pour lemme յոյս (yoys/*espoir*) ; ասես (ases/*tu dis*) ayant pour lemme ասեմ (asem/*dire*) –, la catégorie morphosyntaxique de ce dernier – յոյս est un nom, ասեմ un verbe –, et l'analyse flexionnelle de la forme – յուսով est une forme à l'instrumental, au singulier ; ասես est une forme à l'actif indicatif présent, deuxième personne du singulier –.

À partir de ces données langagières étiquetées, il est alors possible d'éditer différents outils lexicologiques, indispensables à une exploration et une exploitation systématique des textes, à savoir des concordances ou des index lemmatisés[2]. Les illustrations qui suivent fournissent des concordances lemmatisées – monolingue (figure 1) ou bilingue (figure 2) –, ainsi que des

[1] À propos du projet GRE*g*ORI, cfr le site web du projet à l'adresse http://www.uclouvain.be/gregori-project ; la bibliographie complète du projet est disponible sous l'onglet « General bibliography » (tous les liens cités sont actifs au moment de la rédaction de ces lignes) ; cfr aussi les interfaces en ligne du projet sous l'adresse https://www.gregoriproject.com. L'expression « arménien ancien » recouvre ici les différents états de la langue arménienne ancienne, à savoir l'arménien classique, l'arménien hellénisant et le moyen-arménien ; cfr MEILLET, *Elementarbuch*, p. 1-6 (= MEILLET, *Manuel élémentaire*, p. 19-23) ; DE LAMBERTERIE, *Introduction*, p. 235-236. Cet article est la concrétisation d'une communication présentée à la *Conférence internationale "Digital Armenian". Humanités numériques, innovations et nouvelles technologies* (3-5 octobre 2019) organisée à l'Inalco par les laboratoires SeDyL (Inalco-CNRS-IRD), ERTIM (Inalco), l'association ATALA (Paris), l'association Calfa (Paris) et la Société des Études Arméniennes, avec le soutien du laboratoire d'excellence EFL (Paris). Les auteurs tiennent à remercier Chahan Vidal-Gorène et Agnès Ouzounian pour leur implication dans le projet GRE*g*ORI et le partage de leur expertise.

[2] De nombreux exemples de concordances et d'index sont mis à la disposition des chercheurs sous l'onglet « The concordances of the GRE*g*ORI project » du site du projet, cité note 1.

Le Muséon 135 (1-2), 209-241. doi: 10.2143/MUS.135.1.3290656 - Tous droits réservés.

index – index lemmatisé (figure 3) ou index fréquentiel (figure 4) –, générés automatiquement à partir des données présentes dans les corpus. Ces réalisations sont mises à la disposition des chercheurs, sous différents médias, en particulier via des interfaces en ligne, comme l'illustre la figure 5. Fort d'une expérience acquise au fil des années, le projet GRE*g*ORI produit régulièrement des index lemmatisés de textes publiés dans différentes collections d'éditions critiques[3].

[3] En grec : *Basilii Minimi in Gregorii Nazianzeni Orationem XXXVIII Commentarii*, editi a Th. SCHMIDT (*Corpus Christianorum. Series Graeca*, 46. *Corpus Nazianzenum*, 13), Turnhout, 2001 (index grec, p. 121-194); *Severus Sophista Alexandrinus, Progymnasmata quae exstant omnia*, collegit edidit apparatu critico instruxit E. AMATO, cum indice graecitatis a B. KINDT confecto, accedunt Callinici Petraei et Adriani Tyrii sophistarum testimonia et fragmenta necnon incerti auctoris ethopoeia nondum vulgata (*Bibliotheca Scriptorum Graecorum Romanorum Teubneriana*), Berlin – New York, NY, 2009 (index grec, p. 89-130) ; *Procopii Gazaei Epitome in Canticum Canticorum*, edita a J.-M. AUWERS, cum praefatione a J.-M. AUWERS et M.-G. GUÉRARD curata (*Corpus Christianorum. Series Graeca*, 67), Turnhout, 2011 (*Index nominum*, p. 467-469 ; *Index verborum*, p. 470-584) ; *Basilii Minimi in Gregorii Nazianzeni Orationes IV et V Commentarii*, editi a G. RIOUAL, cum indice graecitatis a B. COULIE et B. KINDT confecto (*Corpus Christianorum. Series Graeca*, 90. *Corpus Nazianzenum*, 29), Turnhout, 2019 (index grec, p. 168-264) ; *Sancti Gregorii Nazianzeni Opera. Versio latina*, I. *Epistulae 102-101*, editae ab A. CAPONE, cum indice verborum a B. KINDT et B. COULIE confecto (*Corpus Christianorum. Series Graeca*, 99. *Corpus Nazianzenum*, 32), Turnhout, 2021 (index lemmatisé bilingue grec-latin, p. 25-99). En arménien : *Apocrypha Armeniaca*, I. *Acta Pauli et Theclae, Prodigia Theclae, Martyrium Pauli*, cura et studio V. CALZOLARI (*Corpus Christianorum. Series Apocryphorum*, 20), Turnhout, 2017 (index arménien, p. 706-728) ; *The Genesis Commentary by Step'anos of Siwnik' (Dub.)*, edition, translation and comments by M. STONE, with additional annotations by Sh. EFRATI (*Corpus Scriptorum Christianorum Orientalium*, 695; *Scriptores Armeniaci*, 32), Leuven, 2021 (index arménien, p. 129-220). En géorgien : *Sancti Gregorii Nazianzeni Opera*, *Versio Iberica*, VIII. *Orationes VI, XXIII, XXII*, editae a B. COULIE et Th. OTKHMEZURI, cum indice a B. COULIE et B. KINDT confecto (*Corpus Christianorum. Series Graeca*, 95. *Corpus Nazianzenum*, 30), Turnhout, 2019 (index géorgien, p. 125-235) ; *Vie et conduite des Bienheureux Justes-nus et de notre saint Père Zosime. Trois traductions géorgiennes*, introduction et édition critique par T. PATARIDZE, avec index lemmatisé réalisé en collaboration avec B. COULIE et B. KINDT (*Corpus Scriptorum Christianorum Orientalium*, 686; *Scriptores Iberici*, 25), Leuven, 2020 (index géorgien, p. 33-114). En syriaque : Y. ARZHANOV, *Syriac Sayings of Greek Philosophers. A Study in Syriac Gnomologia with Edition and Translation* (*Corpus Scriptorum Christianorum Orientalium*, 669; *Subsidia*, 138), Leuven, 2019 (index syriaque, p. 327-341); *La versione siriaca della Vita di Giovanni il Misericordioso di Leonzio di Neapolis*, edita da G. VENTURINI (*Corpus Scriptorum Christianorum Orientalium*, 679; *Scriptores Syri*, 263), Leuven, 2020 (index syriaque p. 124-152); A.B. SCHMIDT – B. KINDT, *Eine syrische Amulettrolle mit Beschwörungen für Frauen : Erevan, Matenadaran, rot. syr. 72*, Teil II. *Wortindex*, in *Tracing Written Heritage in a Digital Age*, ed. E.A. ISHAC – Th. CSANÁDY – Th. ZAMMIT LUPI, with an Introduction by R.A. KITCHEN, Wiesbaden, 2021, p. 59-76. En arabe : *Sancti Gregorii Nazianzeni Opera*, *Versio arabica antiqua*, IV. *Orationes XI, XLI (arab. 8, 12)*, editae a J. GRAND'HENRY (*Corpus Christianorum. Series Graeca*, 85. *Corpus Nazianzenum*, 27), Turnhout, 2013 (index arabe, p. 137-252) ; *Sancti Gregorii Nazianzeni Opera*, *Versio arabica antiqua*, V. *Oratio XLII (arab. 14)*, edita a J. GRAND'HENRY (*Corpus Christianorum. Series Graeca*, 98. *Corpus Nazianzenum*, 31), Turnhout, 2020 (index arabe, p. 133-241).

բան

...

GRNA Or. 2 58 9	*Հայցել եւ դատաւորին*	***բանս***	*առ ի չնորհս խաւսել,*
GRNA Or. 2 110 1	*Բայց որ այժմս ինձ*	***բանս***	*ասել դիմեաց, դիտել*
GRNA Or. 2 12 1	*ասիցէ. որպէս իմ է*	***բանս՝***	*եթէ գժուար ընդունելի*
GRNA Or. 2 43 1	*ովագունիւքն զանց արար*	***բանս,***	*զի մի աւելորդիցէ քան*
GRNA Or. 2 49 12	*եթէ երկուս կամ երիս*	***բանս***	*ի բարեպաշտութեանցն*
GRNA Or. 2 2 1	*խոնարհ ճանապարհս*	***բանս***	*մեր դիմեսցէ, յաղագս*
GRNA Or. 2 73 7	*լինիր վաղվաղկոտ ի*	***բանս,***	*նորին Սողոմոնի է ձայն՝*
GRNA Or. 2 20 5	*պատճառաւք մեղաց,*	***բանս***	*չատագովութեան ախտից*
GRNA Or. 2 6 11	*եթէ այլ ոք որ ի*	***բանան***	*են, եւ եի ի մեծամեծ*
GRNA Or. 2 77 13	*հեշտացաւ ինձ Աստուծոյ*	***բանքն***	*իբրեւ զխորխս մեղու*
GRNA Or. 2 45 3	*ասիցէ՝ զզաւրութիւն*	***զբանին***	*կերակուր ոչ բերելով.*
GRNA Or. 2 60 2	*բուռն հարկանէ*	***զբանիցն***	*եւ առ նոյն ինքն գժուարի*
GRNA Or. 2 39 17	*դիւրագոյն առնէ*	***զբանն***	*եւ դիւրընկալ՝ լսողացն*
GRNA Or. 2 46 3	*սրտի մարդոյ*	***զբանն***	*ընդ բազում եւ ընդ դիւրագին*
GRNA Or. 2 46 2	*որք վաճառէն կարեն*	***զբանն***	*ճշմարտութեան եւ խառնեն*

...

Figure 1 : Extrait de la concordance du lemme *բան* (ban/*parole, mot*) dans le Discours 2 de Grégoire de Nazianze

***որդի* ~ N+Com ~ 4**

IRCP 254,6	*ինչ եւ*	***զՈրդին***	*իւր միածին ետ։*
IRCP,FRA,255,6	qu' il a donné son	**Fils**	unique».
IRCP 254,8	*Արդ ունէր սա։ Ե֯։*	***որդի***	*· եւ։ Գ֯։ դուստր։ Ջմի որդին*
IRCP,FRA,255,8	Or, lui [Kostandin] avait 5	**fils**	et 3 filles. L' un de [ses] fils,
IRCP 254,8	*Ջմի*	***որդին***	*Ա[ստուծո]վ Թ[ա]գ[աւոր]եցուցանէ*
IRCP,FRA,255,8	L' un de [ses]	**fils,**	par [la grâce de] Dieu, il le fait roi
IRCP 255,16	*Հանդերձ*	***որդւովք***	*եւ գարմիւք։ · եւ զմերս*
IRCP,FRA,256,16	ainsi que de [ses]	**fils**	et de [sa] famille, et de nous

Figure 2 : Concordance bilingue arménien-français du lemme *որդի* (ordi/*fils, enfant*) dans l'inscription du régent Constantin de Paperōn

աստուած ~ N+Com ~ 51

աստուած ~ 15
5, 12; 5, 14; 5, 24; 5, 26 (2); 5, 27 (3); 5, 28; 5, 29 (2); 5, 31; 5, 32 (2); 5, 33

աստուածոյ ~ 1
5, 32

աստուածոյն ~ 1
5, 40

աստուածս ~ 1
5, 5

աստուածք ~ 3
5, 14; 5, 32 (2)

աստուածքն ~ 2
5, 25; 5, 37

աստուծոյ ~ 18
5, 1 (2); 5, 4 (3); 5, 26; 5, 27 (3); 5, 29; 5, 31; 5, 34; 5, 35 (2); 5, 36; 5, 37; 5, 41; 5, 42

աստուծոյն ~ 1
5, 22

գաստուած ~ 4
5, 7; 5, 26 (2); 5, 35

գաստուածսն ~ 1
5, 32

գաստուծոյ ~ 1
5, 26

գաստուծովն ~ 1
5, 32

յաստուած ~ 1
5, 34

յաստուծոյ ~ 1
5, 2

Figure 3 : Index lemmatisé du lemme *աստուած* (astuac/*dieu*) dans le Discours 5 de Grégoire de Nazianze

...		*սքանչելի*	25
այլ (*այլոյ*)	108	*բարի*	23
ամենայն	98	*այնքան*	21
չար	50	*աստուածային*	18
բազում	49	*փոքր*	18
միայն	47	*միւս*	17
բաց	45	*այնպիսի*	12
քրիստոնեայ	42	*իմաստուն*	12
մեծ	37	*լաւագոյն*	12
որքան	34	*կէս*	12
սակաւ	34	*հասարակաց*	12
առաջին	33	*երեւելի*	11
արժանաւոր	31	*թագաւորական*	11
յոլով	31	*իրաւացի*	11
հակառակ	29	...	

Figure 4 : Index fréquentiel (extrait) des lemmes adjectivaux dans le Discours 4 de Grégoire de Nazianze

En proposant une description des étiquettes flexionnelles (cas, nombres, voix, modes, temps, personnes, etc.) et morphosyntaxiques (noms, adjectifs, pronoms, verbes, adverbes, etc.) utilisées pour le traitement automatique

de l'arménien ancien, cet article s'inscrit directement dans la lignée des travaux du projet et fait suite aux publications antérieures portant sur les principes de lemmatisation suivis pour l'analyse du grec ancien, de l'arabe, du géorgien ou du syriaque[4].

Définir les étiquettes utiles à la description du lexique d'une langue n'est jamais une opération triviale. En arménien, dans l'expression *ըստ առաքելոյ ձայնին* (əst aṙakʿeloy jaynin/*selon la parole de l'apôtre*), la forme *առաքելոյ* (aṙakʿeloy/*de l'apôtre*) sera comprise, par les uns, comme le génitif singulier du nom commun *առաքեալ* (aṙakʿeal/*apôtre*), mais, par d'autres, comme le participe aoriste substantivé du verbe *առաքեմ* (aṙakʿem/*envoyer*), aux mêmes cas. Les deux analyses sont correctes. Elles dépendent de l'approche descriptive adoptée par le lecteur, plutôt synchronique et fonctionnelle dans un cas, ou davantage lexicale et diachronique dans l'autre. Cette situation n'est pas propre à l'arménien. En grec, dans l'expression καὶ Πνεῦμα Κυρίου πεπλήρωκε τὴν οἰκουμένην/*et l'Esprit du Seigneur a rempli le monde*, la forme οἰκουμένην peut être comprise comme un nom (ἡ οἰκουμένη/*le monde*) ou comme un verbe substantivé, au participe (οἰκέω/*habiter*).

Ces interprétations sont intuitives. Dans le cadre d'un traitement automatique, il faut définir des objectifs et, pour les atteindre, arrêter des règles d'analyses concrètes et univoques permettant de décrire le lexique de la langue de manière cohérente. Les choix opérés, s'ils sont explicites et motivés, permettent de proposer un système d'analyse opérationnel sur le long terme, pour différents corpus, même quand plusieurs opérateurs collaborent ou quand différents acteurs se succèdent.

L'approche lexicale du projet GRE*g*ORI vise à déterminer les lexèmes présents dans le vocabulaire des textes étudiés. Il s'agit donc de déterminer le lemme de chaque forme. L'étiquette morphosyntaxique indique alors la catégorie du lemme lui-même, sans tenir compte, à ce stade, des multiples emplois possibles du terme en contexte. L'étiquette morphosyntaxique est dite « prototypique ». Ainsi, en grec, la forme οἰκουμένην reste classée sous le lemme verbal οἰκέω et, en arménien, les formes fléchies apparentées à *առաքեալ* recevront un lemme verbal, *առաքեմ*[5].

[4] Cfr, pour l'analyse du grec ancien, en particulier, COULIE, *Lemmatisation* ; KINDT, *Principes* ; KINDT – PIRARD, *De Nazianze à Ninive* ; KINDT, *Processing Tools* ; pour l'arabe, TUERLINCKX, *Lemmatisation* ; pour le géorgien, COULIE – KINDT – PATARIDZE, *Lemmatisation* ; pour le syriaque, KINDT – HAELEWYCK – SCHMIDT – ATAS, *Concordance bilingue* ; ATAS, *Principles*.

[5] Cfr KINDT, *Principes*, p. 220-225. La forme *Առաքել* (*sic* pour *Առաքեալ*), attestée dans les textes, relève quant à elle d'un lemme anthroponymique *Առաքեալ*, avec majuscule initiale (cfr 2.9).

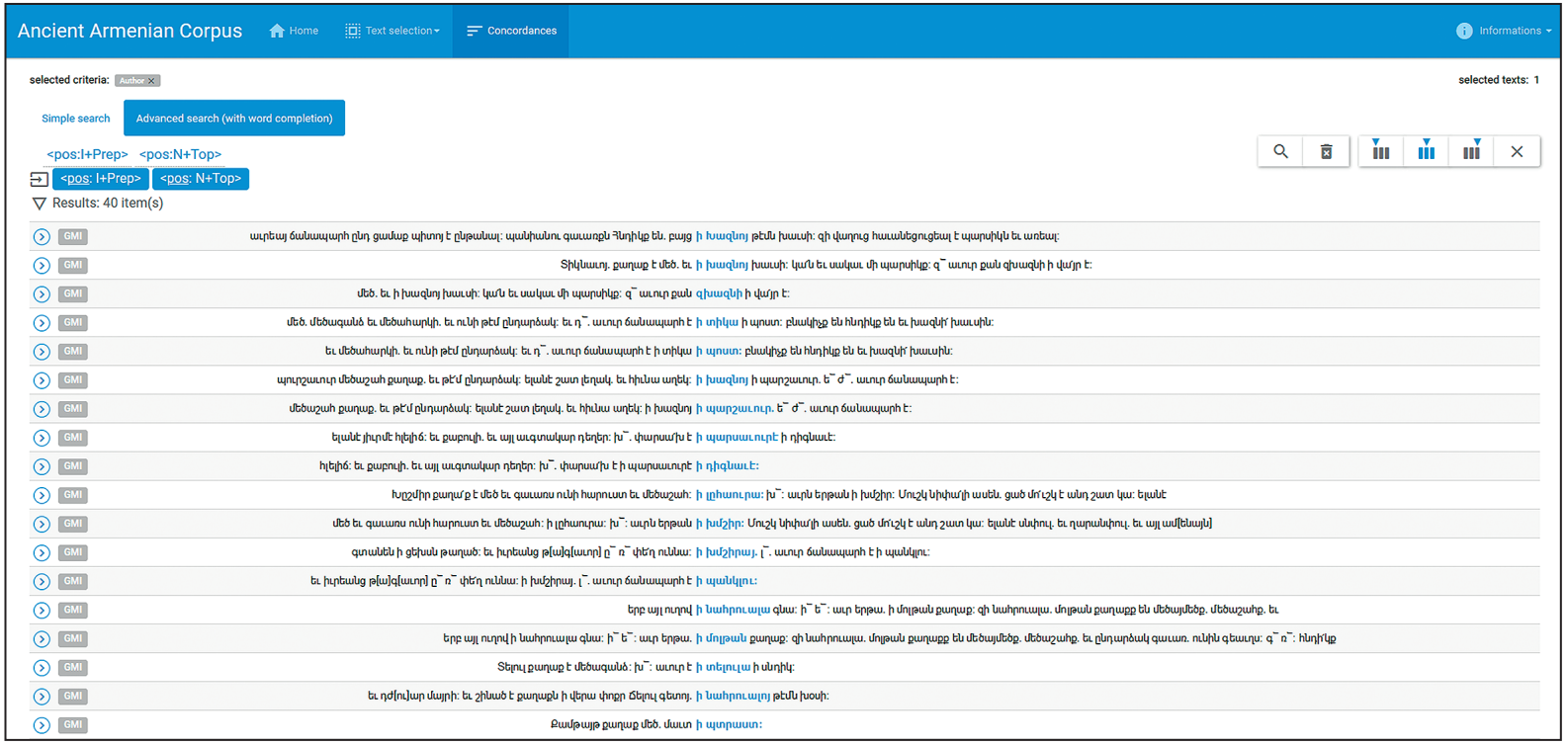

Figure 5 : *Extrait d'une concordance en ligne des séquences constituées d'une préposition (<pos:I+Prep>) suivies d'un lemme toponymique (<pos:N+Top>) dans la* Géographie du monde indien

En ce qui concerne les étiquettes flexionnelles, la situation est plus simple puisque, en ce domaine, un consensus très large existe sur l'identification des informations grammaticales pertinentes: cas, nombres, voix, modes, temps, personnes, etc. Cet article s'attardera donc davantage sur la description des étiquettes morphosyntaxiques, énumérées et illustrées en détail (cfr point 2), que sur les étiquettes flexionnelles (cfr point 3). Mais avant cela, les pages qui suivent identifient rapidement les textes en cours de traitement ou déjà traités (point 1.2.), décrivent les ressources linguistiques mises en œuvre (point 1.3.) et, enfin, précisent le traitement réalisé sur ces textes (point 1.4).

1.2. *Le corpus arménien*

Les travaux sur l'arménien ancien réalisés dans le cadre du projet GRE*g*ORI portent sur différents textes, de taille variable, relevant de différentes époques, de différents genres littéraires et de différents niveaux de langue. Dans la majorité des cas, il s'agit de textes ayant fait l'objet d'une édition critique publiée dans le *Corpus Christianorum*[6], le *Corpus Scriptorum Christianorum Orientalium*[7] ou les volumes de la *Patrologia Orientalis*[8]. Quelques échantillons de la collection des colophons arméniens[9] et

[6] Sur le *Corpus Christianorum*, cfr https://www.corpuschristianorum.org.

[7] Sur le *Corpus Scriptorum Christianorum Orientalium*, cfr http://www.peeters-leuven.be.

[8] Par exemple, *Eznik de Kołb, De Deo*, édition critique du texte arménien, traduction française, notes et tables par L. MARIÈS et Ch. MERCIER (*Patrologia Orientalis*, 28, fasc. 3 et 4), Paris, 1959.

[9] Pour les éditions, cfr A.S. MAT'EVOSYAN (éd.), *Հայերեն ձեռագրերի հիշատակարաններ, ԺԹ դար* (*Նյութեր հայ ժողովրդի պատմության*, 20), [Colophons de manuscrits arméniens, XIII^e^ siècle (*Matériaux d'histoire du peuple arménien*, 20)], Erévan, 1984 ; L.S. XAČ'IKYAN (éd.), *ԺԹ դարի հայերեն ձեռագրերի հիշատակարաններ* (*Նյութեր հայ ժողովրդի պատմության*, 2), [Colophons de manuscrits arméniens du XIV^e^ siècle (*Matériaux d'histoire du peuple arménien*, 2)], Erévan, 1950 ; L.S. XAČ'IKYAN (éd.), *ԺԵ դարի հայերեն ձեռագրերի հիշատակարաններ, Մասն առաջին (1401-1450 թթ.),* (*Նյութեր հայ ժողովրդի պատմության*, 6), [Colophons de manuscrits arméniens du XV^e^ siècle, t. 1 . 1401-1450 (*Matériaux d'histoire du peuple arménien*, 6)], Erévan, 1955 ; L.S. XAČ'IKYAN (éd.), *ԺԵ դարի հայերեն ձեռագրերի հիշատակարաններ, Մասն երկրորդ (1451-1480 թթ.),* (*Նյութեր հայ ժողովրդի պատմության*, 8), [Colophons de manuscrits arméniens du XV^e^ siècle, t. 2. 1451-1480 (*Matériaux d'histoire du peuple arménien*, 8)], Erévan, 1958 ; L.S. XAČ'IKYAN (éd.), *ԺԵ դարի հայերեն ձեռագրերի հիշատակարաններ, Մասն երրորդ (1481-1500 թթ.)* [Colophons de manuscrits arméniens du XV^e^ siècle, t. 3. 1481-1500], Erévan, 1967 ; V. HAKOBYAN – A. HOVHANNISYAN (éd.), *Հայերեն ձեռագրերի ԺԷ դարի հիշատակարաններ, Հատոր Ա (1601-1620 թթ.)*, (*Նյութեր հայ ժողովրդի պատմության*, 14), [Colophons de manuscrits arméniens du XVII^e^ siècle, t. 1. 1601-1620 (*Matériaux d'histoire du peuple arménien*, 14)], Erévan, 1974 ; V. HAKOBYAN – A. HOVHANNISYAN (éd.), *Հայերեն ձեռագրերի ԺԷ դարի հիշատակարաններ, Հատոր Բ (1621-1640 թթ.)*, (*Նյութեր հայ ժողովրդի պատմության*, 15), [Colophons de manuscrits arméniens du XVII^e^ siècle, t. 2. 1621-1640 (*Matériaux d'histoire du peuple arménien*, 15)], Erévan, 1978 ; V. HAKOBYAN (éd.), *Հայերեն ձեռագրերի ԺԷ դարի հիշատակարաններ, Հատոր Գ (1641-1660 թթ.)*, (*Նյութեր հայ ժողովրդի պատմության*, 17), [Colophons de manuscrits arméniens du XVII^e^ siècle, t. 3. 1641-1660 (*Matériaux d'histoire du peuple arménien*, 17)], Erévan, 1984.

de la Bible arménienne (Zohrab)[10] ont aussi été intégrés dans ces travaux. L'examen de ces textes a fourni de nombreux matériaux utiles à la mise au point du système d'analyse de l'arménien présenté dans cette contribution.

Parallèlement, quatre ensembles textuels ont déjà été analysés de manière exhaustive et sont donc munis d'un étiquetage lexical (lemmes) et morphosyntaxique (catégories grammaticales) abouti :

GRNA_HYE : le corpus GRNA_HYE rassemble les textes de la version arménienne des *Discours* de Grégoire de Nazianze déjà publiés dans le *Corpus Christianorum*. Grégoire de Nazianze, mort en 390, est l'auteur de quarante-cinq *Discours*, de plus de deux-cent-quarante lettres, et d'œuvres théologiques et historiques en vers. La version arménienne des *Discours* est anonyme et date de la fin du Ve s. ou du début du VIe s.[11]

Éditions : *Sancti Gregorii Nazianzeni Opera*, *Versio armeniaca*, I. *Orationes II, XII, IX*, ed. B. COULIE (*Corpus Christianorum. Series Graeca*, 28. *Corpus Nazianzenum*, 3), Turnhout, 1994 ; *Sancti Gregorii Nazianzeni Opera*, *Versio armeniaca*, II. *Orationes IV et V*, ed. A. SIRINIAN (*Corpus Christianorum. Series Graeca*, 37. *Corpus Nazianzenum*, 6), Turnhout, 1999; *Sancti Gregorii Nazianzeni Opera*, *Versio armeniaca*, III. *Orationes XXI, VIII*, ed. B. COULIE, *Oratio VII*, ed. A. SIRINIAN (*Corpus Christianorum. Series Graeca*, 38. *Corpus Nazianzenum*, 7), Turnhout, 1999; *Sancti Gregorii Nazianzeni Opera*, *Versio Armeniaca*, IV. *Oratio VI*, ed. C. SANSPEUR (*Corpus Christianorum. Series Graeca*, 61. *Corpus Nazianzenum*, 21), Turnhout, 2007.

THECLA_HYE : le corpus THECLA_HYE rassemble les textes de la tradition arménienne relative à la légende de Thècle et au martyre de Paul, une geste historico-hagiographique produite entre les Ve et XIVe s., en lien avec d'autres sources grecques, latines, syriaques et coptes.

Édition : *Acta Pauli et Theclae*, *Prodigia Theclae*, *Martyrium Pauli*, ed. V. CALZOLARI (*Corpus Christianorum. Series Apocryphorum*, 20. *Apocrypha Armeniaca*, 1), Turnhout, 2017.

GMI : le corpus GMI reprend l'édition diplomatique d'un texte anonyme intitulé *Géographie du monde indien*, daté de 1120 et dressant une liste de villes, de places commerciales et de ports situés en Inde.

Édition : P. BOISSON, *Précis de géographie du monde indien à l'usage des commerçants : édition et traduction annotée*, dans A. MARDIROSSIAN – A. OUZOUNIAN – C. ZUCKERMAN (éd.), *Mélanges Jean-Pierre Mahé* (*Travaux et Mémoires*, 18), Paris, 2014, p. 105-126.

[10] Y. ZŌHRAPEAN, *Աստուածաշունչ մատեան Հին եւ նոր կտակարանաց* (Astuacašunč' matean hin ew nor ktakaranac'/*Livre inspiré par Dieu, contenant l'Ancien et le Nouveau Testament*), Venise, 1805.

[11] Sur la version arménienne des œuvres de Grégoire de Nazianze, cfr LAFONTAINE – COULIE, *Version*, et, plus récemment, sur les différentes versions orientales, HAELEWYCK, *Versions*.

IRCP : le corpus IRCP reprend le texte d'un document épigraphique contenant la dédicace par le régent Constantin du couvent de Papeṙōn, en Cilicie, en 1241.

Édition : M. Goepp – C. Mutafian – A. Ouzounian, *L'inscription du régent Constantin de Papeṙōn (1241). Redécouverte, relecture, remise en contexte historique*, dans *Revue des Études Arméniennes*, 34 (2012), p. 243-287.

Ces quatre ensembles textuels constituent au final un corpus de 67.039 mots-occurrences[12]. Toutes les formes de ce corpus sont identifiées par le lemme qui leur correspond en contexte et chaque lemme est associé à une catégorie morphosyntaxique. Le tableau 1 fournit, pour chacun des textes, le nombre de mots-occurrences, le nombre de formes différentes et le nombre de lemmes.

Corpus	Mots-occurrences	Formes différentes	Lemmes
GRNA_HYE	57.821	14.951	4.759
THECLA_HYE	7.950	2.422	1.1126
GMI	981	434	332
IRCP	287	201	166
Total	67.039	18.008	5.317

Tableau 1 : Le corpus arménien du projet GRE*g*ORI : nombre de mots-occurrences, nombre de formes différentes et nombre de lemmes

Certes, d'un point de vue quantitatif, ce corpus reste bien inférieur, en termes de nombre de mots, aux corpus existant pour les langues modernes en général ou pour des langues classiques et orientales, en particulier. En ce domaine, et pour l'arménien, l'*Eastern Armenian National Corpus* – avec plus de 110 millions de tokens – et le projet ԱՌԱԿ-29 – qui propose notamment une concordance lemmatisée de la Bible arménienne – font figure de pionniers[13].

En revanche, les réalisations du projet GRE*g*ORI se distinguent des initiatives existantes sur différents points :

– tant les outils informatiques (outils de traitement, interfaces de consultation, etc.) que les ressources linguistiques sont développés non seulement pour l'arménien, mais aussi pour le grec et les différentes langues orientales citées – toutes peu dotées en matière de traitement automatique –, ainsi que pour les langues modernes dans lesquelles les textes traités ont été traduits ;

[12] Les concordances lemmatisées de ces textes sont disponibles sur le site du projet, cfr notes 1 et 2.

[13] Cfr, respectivement, http://www.eanc.net et http://arak29.am.

– l'analyse porte sur tous les lexèmes de la langue, en tenant compte des éléments lexicaux inclus dans les crases du grec et dans les formes préfixées, suffixées et agglutinées de l'arabe, du syriaque, du géorgien et de l'arménien ;
– enfin, les principes d'étiquetage sont explicites – décrits dans différentes contributions (cfr note 4) – et donc stables sur le long terme (malgré d'indispensables corrections ou améliorations ponctuelles) et réutilisables pour d'autres corpus, par d'autres opérateurs.

Le traitement des textes ou ensembles de textes mentionnés ci-dessus, déjà lemmatisés ou non, a permis d'appliquer à l'arménien ancien les méthodes d'analyse déjà éprouvées pour le grec et les différentes langues de l'Orient chrétien. Des outils informatiques et des ressources linguistiques sont désormais développés pour l'étude de l'arménien.

1.3. *Les ressources linguistiques*

Le vocabulaire des textes arméniens se partage en deux catégories : les formes dites « simples », d'une part, et les formes polylexicales dites « préfixées » et « suffixées », d'autre part. Le traitement de ces deux types de formes est assuré en utilisant deux types de ressources linguistiques, des dictionnaires pour les premières, et des règles pour les secondes.

1) *Les formes dites « simples »*

À l'instar de la morphologie du grec, celle de la langue arménienne repose sur un système de morphèmes flexionnels. Les formes fléchies rencontrées dans les corpus analysés sont enregistrées dans un fichier informatique qui, pour chaque forme, indique son lemme, sa catégorie morphosyntaxique et son analyse grammaticale, en respectant le formalisme illustré ci-dessous :

Հայր,Հայր.N+Com:As:Ns
Հարք,Հայր.N+Com:Np

Ces deux lignes consignent les éléments suivants :

– en début de ligne, Հայր (hayr) et Հարք (harkʿ) sont des formes simples fléchies ;
– ensuite, après une virgule, Հայր (hayr/*père*), le lemme de ces formes fléchies ;
– après un point, l'indication « N+Com », l'étiquette morphosyntaxique déterminant en l'occurrence la nature nominale du lemme (nom commun) ;

– et enfin, après les doubles points, les abréviations « As », « Ns » et « Np », les étiquettes flexionnelles indiquant l'analyse grammaticale des deux formes fléchies, en l'occurrence « As » et « Ns » pour « accusatif singulier » et « nominatif singulier », pour la première forme, et « Np », « nominatif pluriel », pour la seconde.

Ce formalisme est celui des dictionnaires DELAF, acronyme de « dictionnaires électroniques du Laboratoire d'Automatique Documentaire et Linguistique » ; le « -F » final indique qu'il s'agit d'un dictionnaire de formes simples[14]. Le choix de ce formalisme dans le cadre du projet GRE*g*ORI est ancien. Son intérêt est de proposer, pour chaque unité lexicale décrite, un mode de structuration simple et efficace, mais aussi facilement adaptable aux spécificités d'autres logiciels ou d'autres traitements envisagés.

Il en va de même pour les lemmes relevant des autres catégories grammaticales, comme l'illustrent les exemples suivants pour des verbes (« V »), des adjectifs (« A ») ou des pronoms démonstratifs (« PRO+Dem »):

անկեալ,անկանեմ.V:KJAs:KJNs:KJUs
մեծաբանէին,մեծաբանեմ.V:EÎI3p:MÎI3p
երանելեաց,երանելի.A:Âp:Dp:Gp
սրբոյ,սուրբ.A:Âs:Ds:Gs
այնոսիկ,այն.PRO+Dem:Ap:Up
նոսա,նա (*նա*).PRO+Dem:Ap:Up
etc.

L'ensemble des formes fléchies constitue le *dictionnaire* (des formes simples) du projet GRE*g*ORI. Quand une forme répond à plusieurs analyses possibles, elle fait l'objet de plusieurs entrées dans le dictionnaire, comme l'illustre le cas des formes *այս*, *գան* et *նա* présentées ci-dessous. Les lemmes homographes sont systématiquement distingués les uns des autres par une spécification, ajoutée entre parenthèses, permettant de lever toute ambiguïté.

այս,այս (*սա*).PRO+Dem:As:Ns
այս,այս (*այսոգ*).N+Com:As:Ns
գան,գան (*գանուց*).N+Com:As:Ns
գան,գան (*գալ*).V:EÎP1s
նա,նա (*եւ*).I+Conj

[14] Sur les dictionnaires au format DELAF, cfr SILBERZTEIN, *Dictionnaires* ; COURTOIS – SILBERZTEIN, *Dictionnaires*.

նա,401.NUMA+Car
նա,401e.NUMA+Ord
նա,*նա* (*նա*).PRO+Dem:As:Ns

2) *Les formes dites « préfixées » et « suffixées »*

La morphologie de l'arménien ancien repose également sur l'utilisation d'éléments préfixés (des prépositions et une négation) d'une part, et d'éléments suffixés (articles déterminatifs), d'autre part. Ces préfixes et suffixes se joignent aux formes nominales, verbales et adjectivales, à la manière, *mutatis mutandis*, des éléments caractéristiques des langues agglutinantes. Ils sont considérés comme des unités lexicales à part entière et, à ce titre, ont un lemme. Ces éléments sont :

– les prépositions suffixées : *զ*- (z-), *ց*- (cʿ-), *յ*- (y-), lemmes *զ*[15], *ց* et *ի* ;
– la négation préfixée *չ*- (čʿ-), lemme *ոչ* (očʿ).
– les suffixes démonstratifs : -*ս* (-s), -*զ* (-d), -*ն* (-n), lemmes *ս*, *զ* et *ն*.

Dans ce cas, des règles permettent de représenter les formes préfixées et suffixées en distinguant explicitement les différents éléments lexicaux en présence. La forme *զհայրն* (zhayrn/*le père*) est ainsi analysée comme la combinaison des lemmes « *զ* », « *հայր* » et « *ն* » (z-hayr-n), et ces trois lemmes reçoivent les étiquettes morphosyntaxiques qui leur correspondent, à savoir « I+Prep », « N+Com » et « PRO+Dem » (« préposition », « nom commun » et « pronom démonstratif »).

Il en va de même pour les lemmes d'autres catégories grammaticales, comme l'illustrent les exemples suivants pour les formes *առաքեալն* (aṙakʿealn/*l'apôtre*) et *գերեզմանն* (gerezmann/*le tombeau*) :

formes = *առաքեալ@ն*
lemmes = *առաքեմ@ն*
catégories et analyses = V:KJAs:KJNs:KJUs@PRO+Dem:∅

et

formes = *գերեզմանն@ն*
lemmes = *գերեզման@ն*
catégories et analyses = N+Com:As:Ns@PRO+Dem:∅

Dans ces deux exemples, les arobases (signe typographique « @ ») servent de séparateur entre les informations lexicales, catégorielles et flexionnelles de chaque élément en présence dans la forme. Une telle

[15] Ce lemme regroupe tant les emplois prépositionnels que la *nota accusativi*.

représentation permet d'inventorier tous les éléments lexicaux présents dans un texte et offre aux utilisateurs finaux explorant un corpus la possibilité d'effectuer des recherches sur chacun de ces éléments, même si ces derniers sont « encapsulés » dans une forme préfixée ou suffixée de l'arménien. C'est ce qu'illustrent, par exemple, les concordances (partielles) des occurrences de la négation préfixée չ- (figure 6), du suffixe déterminatif -ը (figure 7), ou d'une séquence constituée de la préposition ը- préfixée à un lemme nominal lui-même suivi d'un lemme verbal (figure 8) dans le corpus THECLA_HYE.

ոչ ~ I+Neg ~ 96
...

ATec 5 16 12	զոր ուսուցանես. քանզի	չէն	սակաւք որ չարախաւսեն զքէն։
ATec 9 27 4	եւ զայս գործեաց Թեկղ, եւ	չէ	յուրաստ զոր ինչ գործեաց.
ATec 16 43 8	եւ բայց ի նմանէ այլ	չիք	Աստուած, կարես կեալ
ATec 11 30 10	սգոյ ի տան իմում, եւ	չիք	ոք որ աւգնէ ինձ, զի ոչ
ATec 2 5 12	կանայս որպէս թէ	չունիցին,	զի նոքա ժառանգեսցեն զերկիր։

...

Figure 6 : Extrait de la concordance des formes présentant le préfixe négatif չ- dans le corpus THECLA_HYE

ը ~ PRO+Dem ~ 33

ATec 4 11 6	կագիցէք, կամ ո՞վ ոք է	այրդ	այդ որ ի ներքս, առ ձեզդ է,
ATec 11 32 6	Ջնջի քաղաքս վասն	անաւրէնութեանդ	զոր արար, կորոյս զամենեսեան
ATec 11 31 8	իմ ես դու, հաւատամ յոր	ապաւինեցայդ	եւ փրկեցեր զիս ի հրոյ անտի.
Prod_I 17 1 2	առաքել առ հչմարիտ	աստուածսիրութիւնդ	ձեր զնկարագիր գեղապանծ
ATec 12 34 10	զանձն քո ի չուրդ, զի չար են	գազանքդ	որ են այդր։"
ATec 3 10 4	զի՞նչ է այդ	գործդ	զոր դու գործես
ATec 11 32 8	կորոյս զամենեսեան	դատաւորդ.	դառն է տեսիլդ զոր տեսանեմք
ATec 9 27 9	եւ ասեն. "Անիրաւ է	դատաստանդ	որով դատիք զԹեկղ։"

...

Figure 7 : Extrait de la concordance des formes présentant le suffixe déterminatif -ը dans le corpus THECLA_HYE

Les dictionnaires et les règles sont deux ressources linguistiques qui se complètent et contribuent à fournir une représentation exhaustive et cohérente du lexique de la langue étudiée.

Le projet GRE*g*ORI a donc pour objectif de fournir des données cohérentes et stables permettant d'étudier le lexique des différentes langues concernées, dans une perspective diachronique (tenant compte par exemple

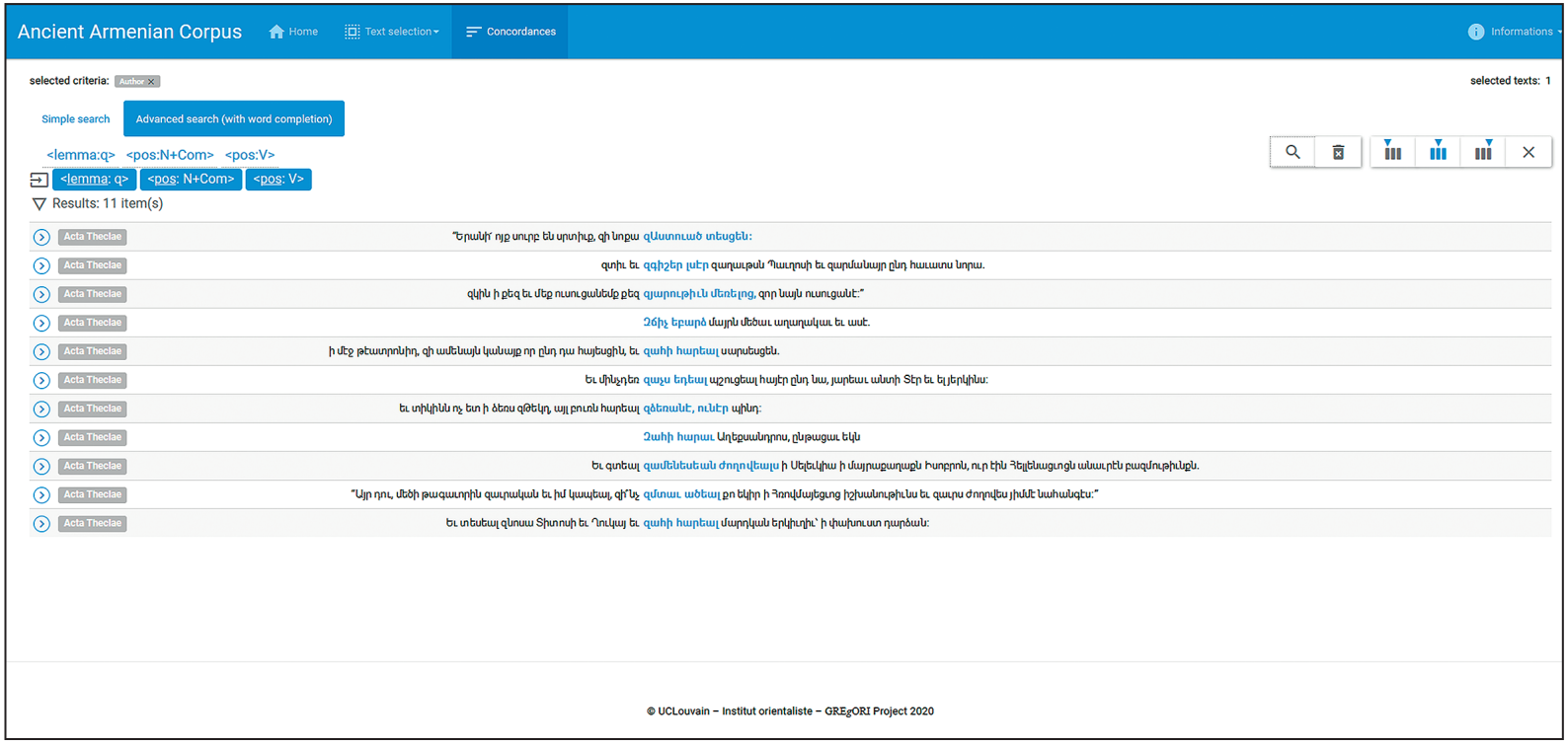

Figure 8 : *Concordance en ligne des séquences constituées de la préposition զ- préfixée à un lemme nominal lui-même suivi d'un lemme verbal dans le corpus THECLA_HYE*

des différents états de la langue au fil de son évolution) et multilingue (tenant compte de la description en parallèle du lexique des différentes versions d'un même texte). Le véritable enjeu est en effet d'analyser le lexique, la syntaxe et le style de chacune des langues traitées de manière à ouvrir la voie à une étude systématique des méthodes de traduction. Les étiquettes utilisées pour l'arménien sont donc semblables à celles utilisées pour les autres langues, en particulier le grec et le géorgien. Pour ces trois langues, les prépositions reçoivent l'unique étiquette « I+Prep », qu'il s'agisse de prépositions autonomes (comme en grec), préfixées (comme en arménien) ou suffixées (comme les postpositions du géorgien), ou même du mot *հանդերձ* [handerj/*avec*] de l'arménien, utilisé comme préposition et placé avant ou après le mot avec lequel il fonctionne. Les suffixes déterminatifs de l'arménien *-ս*, *-դ*, *-ն*, qui servent d'articles définis ou de démonstratifs, ont l'étiquette PRO+Dem, par analogie avec le grec.

Toutes les données lexicales du projet GRE*g*ORI sont issues des textes et ensembles de textes décrits ci-dessus, lemmatisés, en cours de lemmatisation, ou en voie de l'être. Parmi les formes déjà enregistrées dans le dictionnaire, il y a :

– les formes simples attestées dans les corpus GRNA_HYE, THECLA_HYE, GMI et IRCP ;
– des formes simples rencontrées dans les ensembles textuels en cours de traitement ;
– mais aussi des formes simples générées automatiquement, pas encore rencontrées en corpus, mais relevant de lemmes déjà représentés dans les textes traités ou en cours de traitement et qui, à ce titre, relèvent d'un lexique potentiel. Par exemple, si le terme *թագաւոր* (tʻagawor/*roi*) est enregistré dans le dictionnaire sous le lemme *թագաւոր*, le nominatif pluriel équivalent – la forme *թագաւորք* (tʻagaworkʻ/*rois*) – peut sans hésitation être généré afin de compléter d'emblée le paradigme.

Les règles, quant à elles, énumèrent, pour toutes ces formes simples, l'ensemble des combinaisons possibles de ces dernières avec les éléments préfixés et suffixés. Toutes ces combinaisons sont elles aussi générées automatiquement. Ainsi, à partir du lemme *իմաստութիւն* (imastutʻiwn/*sagesse*), sont produites les formes polylexicales *զիմաստութեամբ* (z-imastutʻeamb), *զիմաստութեամբս* (z-imastutʻeamb-s), *զիմաստութեամբդ* (z-imastutʻeamb-d), *զիմաստութեամբն* (z-imastutʻeamb-n), *զիմաստութեամբք* (z-imastutʻeambkʻ), *զիմաստութեամբքս* (z-imastutʻeambkʻ-s), *զիմաստութեամբքդ* (z-imastutʻeambkʻ-d), *զիմաստութեամբքն* (z-imastutʻeambkʻ-n), etc. Ici aussi, de telles formes relèvent du lexique potentiel de la langue. La génération

automatique de formes, pour autant qu'elle soit contrôlée, permet d'accroître avantageusement le nombre de mots reconnus quand un nouveau texte est traité.

Dans leur état actuel, les ressources lexicales arméniennes du projet GRE*g*ORI réunissent 304.231 formes dites « simples », et 760.759 formes dites « préfixées » et « suffixées », soit un total de 1.064.987 formes différentes classées sous 30.303 lemmes. Ces ressources ont récemment été testées sur un ensemble littéraire composé de différents textes en cours de traitement et totalisant 24.011 mots-occurrences. Le nombre des mots-occurrences reconnus (ayant reçu une ou plusieurs analyses possibles) s'élève à 22.044, ce qui représente une couverture lexicale de 91,79%.

1.4. *Le traitement automatique des corpus*

La lemmatisation est l'opération par laquelle chaque forme présente dans un texte traité est reliée à un lemme qui lui correspond dans les ressources linguistiques, ainsi qu'à l'étiquette morphosyntaxique attachée au lemme et à des étiquettes flexionnelles correspondant à l'analyse grammaticale de la forme. Cette opération est réalisée automatiquement à l'aide d'outils informatiques qui comparent le vocabulaire des textes aux données enregistrées dans les ressources linguistiques du projet.

Le tableau 2 fournit un exemple de deux courts extraits du corpus GMI traités de cette manière : *յայս թէնս է ս[ուր]ք առաքեալն Թովմաս վկայեալ։* [...] *Աստի ս[ուր]ք առաքելոյն Թովմայի գերեզմանն ի մալլա՛պն ծովեզեր է* (yays tʿēms ē s[ur]b aṙakʿealn Tʿovmas vkayeal. [...] Asti s[ur]b aṙakʿeloyn Tʿovmayi gerezmann i malla՛pn covezer ē/*dans ce territoire, le saint apôtre Thomas (Tʻovmas) fut martyrisé* [...] *de là au tombeau du saint apôtre Thomas à Mailapur (Mallap) qui se [trouve] sur la côte*) (traduction P. Boisson).

	Références	Formes du texte	Forme des ressources linguistiques	Lemmes	Étiquettes morpho-syntaxiques	Analyses flexionnelles
507	GMI-119-2	*յայս*	*յայս*	*ի@այս (սա)*	I+Prep@ PRO+Dem	As
508	GMI-119-2	*թէնս*	*թէնս*	*թէն@ս*	N+Com@ PRO+Dem	As @∅
509	GMI-119-3	*է*	*է*	*եմ*	V	EÎP3s
510	GMI-119-3	*ս[ուր]ք*	*սուրբ*	*սուրբ*	A	Ns
511	GMI-119-3	*առաքեալն*	*առաքեալն*	*առաքեա@ն*	V@PRO+Dem	KJNs:@∅

	Références	Formes du texte	Forme des ressources linguistiques	Lemmes	Étiquettes morpho-syntaxiques	Analyses flexionnelles
512	GMI-119-3	*Թովմաս*	*թովմաս*	*Թովմաս*	N+Ant	Ns
513	GMI-119-3	*վկայեալ։*	*վկայեալ*	*վկայեմ*	V	KJNs
[...]						
545	GMI-119-6	*Ասի*	*ասի*	*ասի*	I+Adv	
546	GMI-119-6	*ս[ուրբ]ք*	*սուրբ*	*սուրբ*	A	Ns
547	GMI-119-6	*առաքեալոյն*	*առաքեալոյն*	*առաքեա@ն*	V@PRO+Dem	EWGs@∅
548	GMI-119-6	*Թովմայի*	*թովմայի*	*Թովմաս*	N+Ant	Gs
549	GMI-119-6	*գերեզմանն*	*գերեզմանն*	*գերեզման@ն*	N+Com@ PRO+Dem	Ns@∅
550	GMI-119-6	*ի*	*ի*	*ի*	I+Prep	
551	GMI-119-6	*մալա՛պն.*	*մալապն*	*Մալապ@ն*	N+Top@ PRO+Dem	As@∅
552	GMI-119-6	*ծովեզեր*		*ծովեզր*	N+Com	Gs
553	GMI-119-6	*է*	*է*	*եմ*	V	EÎP3s

Tableau 2 : Exemple de texte lemmatisé

Dans la pratique, le texte est chargé dans une base de données dont les différents champs contiennent les informations relatives aux métadonnées (identification du texte ou du corpus, référence au texte dans l'édition, etc.), aux données textuelles (le texte de l'édition) et aux données linguistiques (formes, lemmes, catégories morphosyntaxiques, analyses flexionnelles, etc.).

Dans l'exemple ci-dessus (volontairement simplifié), les premières colonnes réunissent des informations relatives à l'identifiant de chaque mot du texte (par ex., à la première ligne, « 507 ») et à la référence de ces mots dans l'édition utilisée (« GMI-119-2 » ; l'indication « GMI » identifie le texte, « 119 » la page et « 2 » la ligne). Les colonnes qui suivent enregistrent les données proprement linguistiques du texte :

– la forme du mot telle qu'elle apparaît dans le texte de l'édition ;
– la forme du mot telle qu'elle apparaît dans les ressources linguistiques, dépourvue des signes critiques (cfr 510 : *ս[ուր]ք* > *սուրբ*), des marques de ponctuation (cfr 513 : *վկայեալ։* > *վկայեալ*) et des marques d'exclamation (՛), d'interrogation (՞) et d'emphase (՝), ou encore du signe surmontant les déterminants numériques (̅), etc., propres à l'écriture arménienne (cfr 551 : *մալա՛պն.* > *մալապն*) ;

– les analyses correspondant à la forme, à savoir le lemme, la catégorie morphosyntaxique et l'analyse flexionnelle.

Au début du processus d'analyse, les colonnes « Lemmes », « Étiquettes morphosyntaxiques » et « Analyses flexionnelles » sont vides. Le tableau présente le résultat final, après traitement.

Cette méthode fournit des résultats fiables pour l'analyse des formes du texte déjà connues des ressources linguistiques. L'élaboration de ces dernières est largement supervisée par les linguistes et les philologues en charge du traitement des textes. Ainsi, au fil des analyses, les données existantes sont éprouvées sur corpus, et, si nécessaire, complétées ou corrigées. Il y a donc peu de "bruit" (présence de résultats fautifs) lors de leur mise en œuvre.

Par contre, ce *modus operandi* génère du "silence" en fournissant plusieurs résultats pour les formes susceptibles d'avoir des analyses multiples (et donc pas l'analyse univoque qui convient en contexte), et aucun résultat pour les formes encore inconnues des ressources linguistiques. Dans les deux cas, une révision manuelle des résultats obtenus demeure donc, à ce stade, indispensable. Parallèlement, des développements en cours intègrent une méthode d'étiquetage fondée sur des méthodes d'apprentissage machine, en particulier les Recurrent Neural Networks (RNN)[16]. Cette méthode consiste à réaliser un apprentissage supervisé sur des corpus déjà traités et vérifiés, pour générer un modèle d'analyse. Ce modèle est ensuite évalué sur un corpus d'expérimentation afin de l'éprouver et d'en accroître l'efficacité. Il est enfin appliqué à de nouveaux corpus pour fournir une lemmatisation complète proposant une analyse pour toutes les formes du texte, y compris les formes susceptibles de présenter plusieurs analyses ou les formes inconnues des ressources linguistiques. Cette méthode évite le "silence", mais est susceptible de produire du "bruit". Une révision des données reste donc indispensable. Cependant, les expérimentations menées parallèlement sur l'arménien, le géorgien et le syriaque[17], d'abord, et sur le grec[18], ensuite, fournissent des résultats concluants et ouvrent la voie à un traitement des données textuelles exhaustif et plus rapide, diminuant sensiblement les tâches de révision manuelle des données par un expert.

[16] Manjavacas – Kádár – Kestemont, *Improving lemmatization*; Dereza, *Lemmatization*.

[17] Vidal-Gorène – Decours-Perez, *Languages ressources*; Vidal-Gorène – Kindt, *Lemma- and POS-tagging*; Vidal-Gorène – Kindt, *From manuscript to tagged corpora*.

[18] Kindt – Vidal-Gorène – Delle Donne, *Analyse automatique*.

2. *Les étiquettes morphosyntaxiques*

Aucune grammaire ne fournit une liste explicite et exhaustive des catégories morphosyntaxiques de l'arménien ancien. Les catégories abordées dans l'*Altarmenisches Elementarbuch* d'Antoine Meillet sont les suivantes : les adverbes (p. 37-38, §37 ; principalement évoqués dans les paragraphes consacrés à la formation des mots) ; les substantifs et les adjectifs (p. 44-58, §41-60) ; les démonstratifs (p. 59-62, §59-66 ; paragraphes dans lesquels est cité l'article défini) ; les interrogatifs (p. 62-63, §67-68) ; les indéfinis (p. 63-64, §69 ; paragraphes dans lesquels est cité le relatif) ; les possessifs (p. 65, §72) ; les pronoms personnels (p. 66-67, §75-76) ; les nombres (p. 68-70, §84-86; paragraphes dans lesquels on ne tient compte que des noms de nombre ; cfr 2.15, ci-dessous) ; les prépositions (p. 87-90, §101-103 ; dans la morphologie nominale) ; les verbes (p. 91-115, §91-131) ; la négation (p. 124-126, §146-149) ; les particules (p. 132-143, §163-186; paragraphes consacrés en fait aux conjonctions).

L'examen des données textuelles a en revanche permis d'identifier pour l'arménien ancien vingt-sept étiquettes ou combinaisons d'étiquettes, telles que présentées dans le tableau 3. Les intitulés que les étiquettes et les combinaisons d'étiquettes revêtent sont directement inspirés des informations consignées dans les dictionnaires au format DELAF[19]. Elles permettent de décrire le lexique de la langue arménienne, tout en restant proches des étiquettes et combinaisons d'étiquettes utilisées pour la description du lexique des autres langues traitées dans le cadre du projet GRE*g*ORI, afin de faciliter l'interopérabilité des outils informatiques et la mise en relation des données multilingues. Enfin, rappelons qu'elles déterminent la catégorie morphosyntaxique du lemme, indépendamment des multiples emplois rencontrés *in textu*. Ainsi, par exemple, un adjectif substantivé, tel սուրբն (surb-n/*le saint*) conserve un lemme adjectival սուրբ.A, et un participe substantivé, tel առաքեալն (aṙakʿeal-n/*l'apôtre*), reste classé sous le lemme verbal առաքեմ.V (aṙakʿem/*envoyer*)[20].

Des analyses plus fines, tenant davantage compte des emplois des formes en contexte (distinguant participe substantivé, participe épithète ou participe prédicat, etc.), ou faisant intervenir davantage la sémantique (distinction des adverbes de lieu, de manière, de temps ; distinction des conjonctions de coordination ou de subordination etc.), sont dès lors possibles. Mais de telles prolongations ne sont pas inscrites dans les objectifs prioritaires du projet GRE*g*ORI.

[19] Cfr ci-dessus, 1.3.
[20] Cfr ci-dessus, 1.1.

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

TABLEAU 3 : liste des étiquettes et combinaisons d'étiquettes morphosyntaxiques

2.1. *A = Adjectif*

Les adjectifs reçoivent l'étiquette A : *մեծ* (mec/*grand*), *աստուածա-շունչ* (astuacašunčʿ/*inspiré par Dieu*). Relèvent également de lemmes adjectivaux :

– les adjectifs verbaux en -*լի* (-li) formés sur des infinitifs : *սիրեմ* (sirem/*aimer*) > *սիրելի* (sireli/*aimable*), *զարհուրիմ* (zarhurim/*redouter*) > *զարհուրելի* (zarhureli/*redoutable*) ;
– les adjectifs substantivés : *սուրբն* (surbn/*le saint*) ;
– les adjectifs en -*ացի*/-*եցի* utilisés comme ethniques ou désignant les adeptes d'une hérésie : *Հռովմայեցի* (Hṙovmayecʿi/*romain*), *Նիկողիացի* (Nikołiacʿi/*Nicolaïte*);
– les adjectifs employés comme adverbes : *ազատ* (azat/*noble*, *noblement*), *ակներեւ* (aknerew/*évident*, *évidemment*) ;
– les adjectifs employés comme prépositions : *հուպ* (hup/*près*), *մերձ* (merj/*près*) ;
– les adjectifs employés comme interjections : *եղուկ* (ełuk/*hélas*) ; *երանի* (erani/*bienheureux*).

2.2. *I+Adv = Mot invariable – Adverbe*

Les adverbes reçoivent l'étiquette I+Adv : *անդ* (and/*là*), *այժմ* (ayžm/*maintenant*), *աշխատասիրաբար* (ašxatasirabar/*avec application*). À ce niveau, il n'y a pas de distinction entre adverbes de lieu, de temps, de manière, etc.

2.3. *I+AdvPr = Mot invariable – Adverbe prépositionnel*

Les adverbes prépositionnels – adverbes évoluant vers un emploi prépositionnel régissant un cas – reçoivent l'étiquette I+AdvPr : *արտաքոյ* (artakʿoy/*en dehors*, *par-dessus, hors de*), *յանդիման* (yandiman/*devant*, *en présence de*).

2.4. *I+Conj = Mot invariable – Conjonction*

Les conjonctions reçoivent l'étiquette I+Conj[21] : *այլ* (*եւ*) (ayl [ew]/*mais*), *արդ* (ard/*or*, *donc*, *maintenant*), *բայց* (bayc ʿ/*mais*), *եթե* (etʿe/*si*)[22], *եւ* (ew/*et*), *զի* (zi/*car, afin que, que* [emploi causal, final, complétif]), *քանզի* (kʿanzi/*car*), *թեպէտ* (tʿepēt/*bien que*)[23], *իբր* (ibr/*comme*), *իբրեւ* (ibrew/*comme*), *իբրու* (ibru/*puisque*), *իսկ* (isk/*mais*), *կամ* (kam/*ou*), *մինչդեռ* (minčʿdeṙ/*pendant que*), *մինչեւ* (minčʿew/*jusqu'à ce que*), *յորժամ* (yoržam/*lorsque*), *նա* (*եւ*) (na [ew]/*cependant*), *որովհետեւ* (orovhetew/*parce que*), *որչափ* (orčʿapʿ/*autant que*), *որպէս* (orpēs/*de même que*), *ուրեմն* (uremn/

[21] Dans la majorité des cas, ces énumérations ne constituent pas un inventaire exhaustif des lemmes relevant de la catégorie décrite ; ces listes restent donc ouvertes et l'examen de nouveaux textes justifiera de les préciser et de les compléter.

[22] Ce lemme regroupe les formes *թե* et *եթե*, ainsi que *թէ* et *եթէ*.

[23] Ce lemme regroupe les formes *թեպէտ* et *թէպէտ*.

donc), ու (u/*et*), քան (kʿan/*que*), etc. À ce niveau, il n'y a pas de distinction entre les conjonctions de coordination et de subordination.

2.5. *I+Intj = Mot invariable – Interjection*

Les interjections reçoivent l'étiquette I+Intj : աշ (աշ) (ah [ah]/*ah*)[24], աշա (aha/*voici*), աշաւադիկ (ahawadik/*voilà*), աշաւանիկ (ahawanik/*voilà*, *voici*), աշաւասիկ (ahawasik/*voilà*, *voici*), ադէ (ałē/*hé*, *hé bien*), այո (ayo/*oui*), ամէն (աշ) (amēn [ah]/*amen*)[25], աւադ (awał/*hélas*), աւն (awn/*allons*, *hé bien*), աւչ (awš/*oh*)[26], ափսոս (apʿsos/*hélas*), բաբէ (babē/*waouh*), է (ē/*hé*), հայ (hay/*oui*), ով (ov/*oh*)[27], վայ (vay/*malheur*, *vae*)[28], վաշ (vaš/*bravo*), վիշ (viš/[dénote un soupir]), տի (նա) (ti [na]/*allons*), etc.

2.6. *I+Neg = Mot invariable – Négation*

Les négations reçoivent l'étiquette I+Neg. L'arménien connaît deux négations : մի (ոչ) (mi [očʿ])[29] exprimant la défense – en principe pourvue d'un šešt (մի՛) –, et ոչ (očʿ), la négation simple. Ce dernier lemme regroupe aussi la forme préfixée չ- (čʿ-), comme dans չեւ (čʿew/*non plus*) et չիք (čʿikʿ/*il n'y a pas*).

2.7. *I+Part = Mot invariable – Particule*

Les particules reçoivent l'étiquette I+Part. À ce stade des analyses, cette catégorie ne regroupe que les deux particules verbales կու et տի attestées dans des textes en moyen-arménien et classées respectivement sous les lemmes կու (կը) (ku [kə])[30] et տի (պիտի) (ti [piti])[31].

2.8. *I+Prep = Mot invariable – Préposition*

Les prépositions reçoivent l'étiquette I+Prep : առ (aṙ/*à cause de*, *auprès de*, *pour*, *au bord de*), զ (z/*au sujet de*, *contre*, *autour*, *au-delà de*)[32], ընդ

[24] Ce lemme regroupe les formes ա, այ, աշ, ափ et հա. La spécification permet de distinguer le lemme de l'interjection du lemme nominal աշ (աշիւղ) (ah [ahicʿ]/*la peur*).

[25] Ce lemme regroupe les formes ամեն, ամէն et ամին. La spécification « (աշ) » permet de distinguer le lemme de l'interjection du lemme adjectival ամէն (ամենիւղ) (amēn [amēnicʿ]/*tout*, *tous*, *chaque*). La majuscule initiale permet de distinguer ces formes de l'anthroponyme Ամէն (Amēn), cfr 2.9.

[26] Ce lemme regroupe les formes աւչ et աւխ.

[27] Ce lemme regroupe les formes ո, ոշ et ով.

[28] Ce lemme regroupe les formes վայ, վաշ, վա et վախ.

[29] La spécification permet de distinguer la négation du lemme numérique մի (միոյ) (mi [mioǰ]/*un*).

[30] Ce lemme regroupe les formes կը, կի, կ՚ et կ-. La spécification permet de distinguer la particule du lemme nominal կու (կոյ) (ku [koy]/*fiente*, *fumier*).

[31] La spécification permet de distinguer la particule du lemme nominal տի (տա) (ti [am]/*année*) et de l'interjection տի (նա) (ti [na]/*allons!*, *courage!*).

[32] Ce lemme regroupe tant les emplois prépositionnels que la *nota accusativi*.

(ənd/*à la place de*, *sur*, *par*, *dans*, *avec*, *sous*, *entre*), ըստ (əst/*d'après*, *en raison de*, *au-delà*), ի (i/*dans*, *depuis*, *en vue de, etc.*)[33], հանդերձ (handerj/*avec*)[34], ց (cʻ/*à*), *etc.*

Quand ces différents éléments sont rassemblés dans une locution prépositionnelle, chaque élément constitutif de cette locution est décrit séparément et conserve donc le lemme et l'étiquette morphosyntaxique qui lui sont propres, même si certains linguistes et philologues considèrent ces locutions comme des unités polylexicales à analyser dans leur ensemble : առ ի (aṙ i/*afin de*) ; ի վերայ (i veray/*sur*) ; ի ներքոյ (i nerkʻoy/*au-dessous*) ; ի ձեռն (i jeṙn/*via*). Puisque les interfaces en ligne du projet GRE*g*ORI permettent de rechercher des séquences de mots (cfr figure 5 et 8, ci-dessus), il est possible d'extraire d'un corpus les séquences constituées de deux lemmes successifs utilisées comme locutions prépositionnelles, comme l'illustre la figure 9.

2.9. *N+Ant = Nom propre anthroponymique*

Les noms propres anthroponymiques reçoivent l'étiquette N+Ant : Անդրէաս (Andrēas/*André*), Զենովն (Zenovn/*Zénon*) Սաբեղլիոս (Sabełlios/*Sabellius*). Le lemme Քրիստոս (Kʻristos/*Christ*) est également catégorisé comme nom propre.

Les emplois anthroponymiques de mots relevant originellement d'une autre catégorie justifient toujours la création d'un lemme distinct, étiqueté N+Ant : p.ex. des personnes dénommées Առաքեալ (*Aṙakʻeal*) (cfr note 5) dérivé du verbe առաքեմն (aṙakʻem/*envoyer*), Կարապետ (*Karapet*) dérivé du nom կարապետ (karapet/*précurseur*), Աւագ (Awag) dérivé de l'adjectif աւագ (awag/*grand*), Ամէն (Amēn) dérivé de l'interjection ամէն classée sous le lemme ամէն (աՀ) (amēn [ah]/*amen*).

2.10. *N+Com = Nom commun*

Les noms communs reçoivent l'étiquette N+Com : ամ (am/*année*), որդի (ordi/*fils*), տուն (tun/*maison*).

Les lemmes աստուած (astuac/*dieu*) et է (ē/[ce qui existe] *l'Être*) sont considérés comme noms communs, indépendamment de la casse de leur initiale ou du sens qu'ils revêtent en contexte.

Une forme nominale fléchie employée comme préposition ne justifie pas la création d'un nouveau lemme. La forme աղագաւ (ałagaw/*à cause de*), instrumental de աղագ (ałag/*raison*, *occasion*), conserve un lemme nominal.

[33] Ce lemme regroupe aussi la forme préfixée յ- (y-).

[34] Ce lemme conserve son étiquette I+Prep, qu'il soit utilisé comme préposition ou comme postposition.

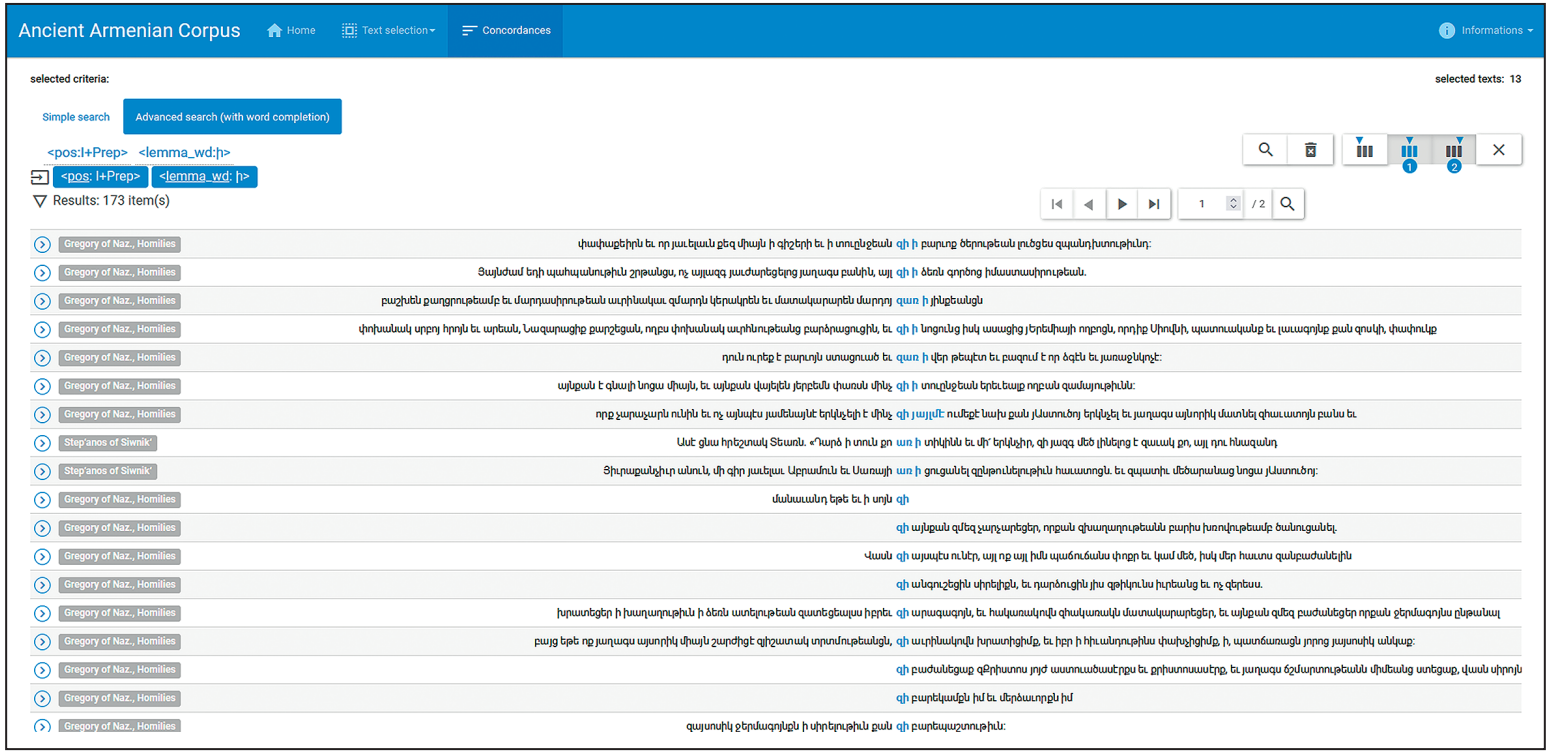

Figure 9 : *Extrait d'une concordance en ligne des séquences constituées de deux prépositions successives (<pos:I+Prep><pos:I+Prep>) dans le corpus THECLA_HYE*

2.11. *N+Let = Nom d'une lettre*

Les lettres citées *in textu* par un caractère alphabétique reçoivent un lemme propre caractérisé par l'étiquette N+Let : *ա s.l. այբ*.N+Let (ayb), *բ s.l. բեն*.N+Let (ben), *գ s.l. գիմ*.N+Let (gim), etc.

2.12. *N+Pat = Nom propre patronymique*

Les noms propres patronymiques (noms de famille, noms de dynastie, etc.) reçoivent l'étiquette N+Pat : *Մամիկոնեան* (Mamikonean/*Mamikonian*), *Ռուբինեանց* (Ṙubineancʿ/*Roupénide[s]*), *Մինասեանց* (Minasencʿ/*Minasencʿ*) comme dans la phrase *զբնտրեալ փիլիսոփայն զՄինասեանց Թումայն* (zəntreal pʿilisopʿayn zMinasencʿ Tʿumayn/*l'excellent philosophe Tʿumay Minasencʿ*).

2.13. *N+Prop = Nom propre*

Les noms propres qui ne sont ni des anthroponymes (N+Ant), ni des patronymes (N+Pat), ni des toponymes (N+Top), reçoivent l'étiquette N+Prop : *Ակադեմիա* (Akademia/*l'Académie*).

2.14. *N+Top = Nom propre toponymique*

Les noms propres toponymiques reçoivent l'étiquette N+Top : *Ասիա* (Asia/*Asie*), *Բեթղեհէմ* (Betʿłehēm/*Bethléem*), *Իկոնիոն* (Ikonion/*Iconium*).

2.15. *NUM et NUMA = Déterminants numériques (écrits en toutes lettres ou écrits selon le système alpha-numérique)*

Deux types de déterminants numériques doivent être distingués :

1) les déterminants numériques écrits en toutes lettres : *երկու* (erku/*deux*), *երկրորդ* (erkrord/*deuxième*), *տասն* (*tasn*/dix), *տասներորդ* (*tasnerord*/dixième). Ces déterminants sont considérés comme des mots et reçoivent l'étiquette NUM.

Quand ces différents éléments sont rassemblés dans une locution, chaque élément constitutif de cette locution est décrit séparément et conserve donc le lemme et l'étiquette morphosyntaxique qui lui sont propres. Dans l'expression *եւթն եւ տասն* (ewtʿn ew tasn/*dix-sept*), les différents éléments sont respectivement classés sous *եւթն*.NUM+Car, *եւ*.I+Conj et *տասն*. NUM+Car. Ici encore, comme dans le cas des locutions prépositives, les interfaces en ligne du projet GRE*g*ORI permettent de rechercher des séquences de mots et donc de reconstituer ces locutions numériques dans l'interrogation et dans les réponses.

2) les déterminants numériques écrits dans le système alphanumérique : բ (*b :* 2 ou 2e), խ (*x :* 40 ou 40e). Ces déterminants sont considérés comme des chiffres ou des nombres et reçoivent l'étiquette NUMA. Dans le texte des éditions, ces déterminants sont souvent, mais pas systématiquement, surmontés du signe « ¯ » (բ̄, խ̄) ou notés entre doubles-points « : » (« :ԽԱ: »).

Les lemmes étiquetés NUM et NUMA sont aussi des déterminants cardinaux (NUM+Car et NUMA+Car) ou ordinaux (NUM+Ord et NUMA+Ord). Les lemmes երկու et երկրորդ reçoivent respectivement les étiquettes NUM+Car et NUM+Ord. La forme ա utilisée comme déterminant numérique sera classée soit sous le lemme « 1.NUMA+Car » soit sous « 1er.NUMA+Ord »[35], la forme բ, sous « 2.NUMA+Car » ou « 2e.NUMA+Ord » et la forme գ sous « 3.NUMA+Car » ou « 3e.NUMA+Ord », etc.

Pour ces déterminants numériques écrits dans le système alphanumérique, les éditeurs (à l'instar des copistes) peuvent opter pour une graphie faisant apparaître un morphème flexionnel correspondant au cas du déterminant ou un suffixe qui en précise le caractère ordinal : la forme ճ՚Ծ-ից (150-icʿ/150) reçoit l'étiquette NUMA+Car; la forme ժ-երորդ (10-erord/*dixième*) reçoit l'étiquette NUMA+Ord.

Les mots désignant les dizaines et les centaines reçoivent l'étiquette NUM+Car : քսան (kʿsan/*vingt*), երեսուն (eresun/*trente*), հարիւր (hariwr/*cent*), երկերիւր (erkeriwr/*deux cents*), երեքհարիւր (erekʿhariwr/*trois cents*), հազար (hazar/*mille*), etc. En revanche, բեւր (bewr/*myriade*) conserve un lemme nominal (N+Com).

2.16. *PRO+Dem = Pronom démonstratif*

Les pronoms démonstratifs reçoivent l'étiquette PRO+Dem, qu'ils soient utilisés en fonction pronominale ou adjectivale : սա (sa/*lui* [proche de moi]), դա (da/*lui* [proche de toi]), նա (na/*lui*), այս (ays/*celui-ci*, *ce …-ci*), այդ (ayd/*celui-là*, *ce …-là*), այն (ayn/*celui-là* [là-bas], *ce …-là* [là-bas]), սոյն (soyn/*même*), դոյն (doyn/*même*), նոյն (noyn/*même*). Cette étiquette regroupe aussi les suffixes déterminatifs -ս (-s), -դ (-d) et -ն (-n).

2.17. *PRO+Ind = Pronom indéfini*

Les pronoms indéfinis reçoivent l'étiquette PRO+Ind : երկաքան-չիւր (erkakʿančʿiwr/*chacun des deux*), ոք (okʿ/*quelqu'un*), ոմն (omn/*quelqu'un*), ինչ (inčʿ/*quelque chose*), իմն (imn/*quelque chose*), իւրաքան-չիւր (iwrakʿančʿiwr/*chacun*) – que certaines grammaires catégorisent comme

[35] Sans oublier que la forme ա relève aussi du lemme աՀ (աՀ).I+Intj, comme expliqué en 2.5, et du lemme այբ.N+Let (ayb), comme expliqué en 2.11.

pronom distributif –, իք (ikʿ/*quelque chose*). Le mot ամենայն (amenayn/*tout*) et les collectifs ամենեքին/ամենեքեան (amenekʿin/amenekʿean/*tous ensemble*), բոլորեքին/բոլորեքեան (bolorekʿin/bolorekʿean/*tous ensemble*) sont également analysés comme des PRO+Ind.

2.18. *PRO+Int = Pronom interrogatif*

Les pronoms interrogatifs reçoivent l'étiquette PRO+Int : ո՞ (oˀ/*qui ?*)[36], ո՞ր (oˀr/*quel ?*), ի՞նչ (iˀnčʿ/*que*, *quel*, *lequel?*), ի՞ (iˀ/*quoi ?*), etc. La plupart des grammaires de l'arménien ancien donnent les formes զի՞նչ et զի՞ pour l'interrogatif; il s'agit formellement de l'interrogatif précédé de la préposition զ-, et des textes tardifs – rencontrés dans les analyses du corpus GREgORI – présentent la forme simple ի՞նչ, qui est dès lors retenue comme lemme; par analogie, la forme ի՞ est également retenue comme lemme.

La présence du signe interrogatif « ՞ » dans l'intitulé du lemme permet de distinguer les pronoms interrogatifs de l'indéfini (cfr 2.17) et du relatif (cfr 2.23).

2.19. *PRO+Per[1,2][s,p] = Pronom personnel*

Les pronoms personnels reçoivent les étiquettes PRO+Per1s, PRO+Per1p, PRO+Per2s et PRO+Per2p. : ես (es/*je*), դու (du/*tu*), մեք (mekʿ/*nous*), դուք (dukʿ/*vous*).

2.20. *PRO+Pos[1,2][s,p] = Pronom possessif*

Les pronoms possessifs reçoivent les étiquettes PRO+Pos1s, PRO+Pos2s, PRO+Pos1p et PRO+Pos2p, qu'ils soient utilisés en fonction adjectivale ou en fonction pronominale : իմ (im/*mon*, *le mien*), քո (kʿo/*ton*, *le tien*), մեր (mer/*notre*, *le nôtre*), ձեր (jer/*votre*, *le vôtre*). Le pronom possessif de la troisième personne est rendu par les pronoms démonstratifs (cfr 2.16) et réfléchis (cfr 2.22).

2.21. *PRO+Rec = Pronom réciproque*

Les pronoms réciproques reçoivent l'étiquette PRO+Rec : իրեար (irear/*l'un l'autre*) et միմեանս (mimeans/*l'un l'autre*).

2.22. *PRO+Ref = Pronom réfléchi*

Les pronoms réfléchis reçoivent l'étiquette PRO+Ref : իւր (iwr/*soi-même*) et ինքն (inkʿn/*lui-même*).

[36] Ce lemme regroupe les formes ո՞ et ո՞վ.

2.23. *PRO+Rel = Pronom relatif*

Le pronom relatif *որ* (or/*qui*) reçoit l'étiquette PRO+Rel.

2.24. *V = Verbe*

Les verbes reçoivent l'étiquette V : *բերեմ* (berem/*porter*), *լամ* (lam/*pleurer*), *հեղում* (hełum/*puiser*, *verser*).

Les formes participiales en *-եալ*, même substantivées, conservent un lemme verbal : *առաքեալ* (aṙakʿeal/*apôtre*) est classé sous *առաքեմ* (aṙakʿem/*envoyer*).

Les adjectifs verbaux en *-ոյ* conservent un lemme verbal : *սրբելոյ* (srbelocʿ/*qui doit être sanctifié*) est classé sous *սրբեմ* (srbem/*sanctifier*), *գալոյ* (galocʿ/*qui doit venir*) est classé sous *գամ* (gam/*venir*).

Les formes supplétives servant d'aoriste sont classées sous le lemme au présent dont elles complètent le paradigme : *գամ* (gam/*je viens*) recouvre aussi les formes du type *եկի* (eki) ; *երթամ* (ertʿam/*je vais*) pour les formes du type *չոգայ* (čʿogay) ; *ընպեմ* (əmpem/*je bois*) pour les formes du type *արբի* (arbi) ; *ընդունիմ* (əndunim/*je reçois*) pour les formes du type *ընկալայ* (ənkalay) ; *ունիմ* (unim/*j'ai*, *je tiens*) pour les formes du type *կալայ* (kalay) ; *ուտեմ* (utem/*je mange*) pour les formes du type *կերայ* (keray), etc.

Pour les verbes défectifs (rares en arménien), une forme de la troisième personne du singulier du présent de l'indicatif ou, si cette dernière n'existe pas, la forme de l'impératif présent servent de lemme : *գոգ* (*ասեմ*) (gog [asem]/*dire*)[37] ; *գոյ* (goy/*il existe*), etc.

Les formes verbales périphrastiques, composées d'un participe et d'un verbe *être*/*devenir*, du type *եկեալ էր* (ekeal ēr/*il était venu*) sont analysées séparément, comme deux formes distinctes, puisque les interfaces en ligne du projet GRE*g*ORI permettent de rechercher des séquences de mots et donc de reconstituer ces formes périphrastiques dans l'interrogation et dans les réponses.

Enfin, l'analyse des formes indique la voix – active, moyenne, passive – lorsque cette analyse est pertinente, c'est-à-dire pour les formes personnelles du verbe. La voix n'est pas indiquée pour les infinitifs et les participes. En arménien ancien, les infinitifs ont à la fois un sens actif, moyen ou passif; les infinitifs passifs tardifs en *-իլ* de verbes actifs en *-եմ* se reconnaissent d'eux-mêmes comme des passifs. Les participes ont également des sens actifs, moyens et passifs, qui ne sont pas distingués dans l'étiquetage.

[37] La spécification permet de distinguer le lemme verbal du nom *գոգ* (*գոգոց*) (gog [gogocʿ]/*sein*, *golfe*).

3. *Étiquettes flexionnelles*

Les étiquettes flexionnelles sont décrites dans le tableau 4, ci-dessous :

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

TABLEAU 4 : Liste des étiquettes flexionnelles

4. *Conclusions et perspectives*

Forts des acquis décrits dans les pages qui précèdent, les travaux du projet GRE*g*ORI, en ce qui concerne l'arménien ancien, s'orientent désormais, et naturellement, selon trois axes complémentaires : 1) accroissement

des données ; 2) amélioration des traitements ; 3) diffusion des corpus lemmatisés.

En ce qui concerne l'accroissement des données, deux projets importants sont en cours. Le premier, mené en collaboration avec la maison d'édition Peeters Publishers (Leuven, Belgique), porte sur les textes publiés dans la série des *Scriptores Armeniaci* du *Corpus Scriptorum Christianorum Orientalium* (CSCO_HYE ; trente-cinq volumes parus depuis 1953). Pour les textes des publications les plus récentes (à partir de 1998), le corpus a été constitué à partir de fichiers transmis par l'éditeur. Les textes issus de publications antérieures à cette date ont fait l'objet d'un traitement par OCR réalisé avec les moyens techniques du projet Calfa (Paris, France)[38]. Tant les textes originaux (en arménien) que les traductions dans une langue moderne (en anglais, français ou italien) sont en cours de lemmatisation. Les textes sources (en arménien) seront de plus alignés sur les textes cibles (leur traduction). Le corpus arménien CSCO_HYE totalise 508.898 mots-occurrences et 51.562 formes différentes.

Un second projet, mené en collaboration avec le Matenadaran d'Erevan[39] (Arménie) et le projet Calfa, vise à produire le corpus lemmatisé des textes publiés dans la collection du *Matenagirkʿ Hayocʿ* (MH ; vingt-et-un volumes parus depuis 2001). La taille de ce corpus MH est estimée à 7.508.000 mots-occurrences.

Enfin, deux projets sont en cours à l'Institut orientaliste de l'Université catholique de Louvain. Le premier porte sur l'étude systématique du vocabulaire des colophons arméniens[40], le second sur les versions arméniennes de l'*Apocalypse* de Jean publiées par Fr. Murat et F.C. Conybeare[41].

[38] Cfr https://calfa.fr; VIDAL-GORÈNE – DECOURS-PEREZ, *Languages resources*.

[39] Cfr http://www.matenadaran.am.

[40] Cfr https://uclouvain.be/fr/instituts-recherche/incal/ciol/colophons-of-armenian-manuscripts.html. L'analyse lexicale des colophons arméniens est menée sous la responsabilité scientifique d'Emmanuel Van Elverdinghe; cfr E. VAN ELVERDINGHE, *Recurrent Pattern Modelling in a Corpus of Armenian Manuscript Colophons*, dans *Journal of Data Mining & Digital Humanities*, January 11, 2018, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages; E. VAN ELVERDINGHE, *Modèles et copies. Étude d'une formule des colophons de manuscrits arméniens (VIII^e^-XIX^e^ siècles),* (*Corpus Scriptorum Christianorum Orientalium*, 698; *Subsidia*, 146), Louvain, 2022. Pour les éditions, cfr note 9.

[41] L'étude du lexique de ces textes s'inscrit dans le projet d'*Editio Critica Maior* de l'*Apocalypse* mené sous l'égide de l'*Institut für Septuaginta- und biblische Textforschung* (Wuppertal, Allemagne ; cfr https://apokalypse.isbtf.de). Pour les éditions, cfr Fr. MURAT, *Yaytnutʿeann Yovhannu hin hay tʿargmanutʿiwn*, Jérusalem, 1905-1911, p. 3-76 ; F.C. CONYBEARE, *The Armenian Version of* Revelation *and Cyril of Alexandria's* Scholia on the Incarnation *and* Epistle on Easter (*Text and Translation Society*, 5), Londres, 1907, p. 1-32.

En matière de traitement, face à de tels volumes de texte, il sera indispensable de poursuivre l'enrichissement des ressources linguistiques (telles que décrites ci-dessus ; cfr 1.3) et de développer aussi des modes d'analyse relevant de la technologie des RNN (comme évoqué ci-dessus, cfr 1.4).

Enfin, des interfaces d'exploration et d'interrogation des textes sont désormais accessibles aux chercheurs (cfr note 1 et figures 5, 8 et 9, ci-dessus) ; les corpus sont ainsi consultables sur l'écran d'un ordinateur personnel. Il s'agit, en fait, de négocier un virage numérique qui permettra de valoriser, de pérenniser et de diffuser les textes écrits en arménien ancien – comme le projet le fait pour les textes grecs et les textes écrits dans les autres langues de l'Orient chrétien – et d'assurer ainsi la transmission de l'héritage culturel qui en découle.

Bibliographie

ATAS, *Principles* = N. ATAS, *Principles of Syriac Lemmatisation. Summary version*. UCLouvain – Institut orientaliste – GRE*g*ORI Project 2021 (document accessible en ligne sur les interfaces du projet GRE*g*ORI, cfr note 1).

COULIE – KINDT – PATARIDZE, *Lemmatisation* = B. COULIE – B. KINDT – T. PATARIDZE, *Lemmatisation automatique des sources en géorgien ancien*, dans *Le Muséon*, 126/1-2 (2013), p. 161-201.

COULIE, *Lemmatisation* = B. COULIE, *La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs*, in *Byzantion*, 66 (1996), p. 35-54.

COURTOIS – SILBERZTEIN, *Dictionnaires* = B. COURTOIS – M. SILBERZTEIN, *Dictionnaires électroniques du français*, dans *Langue Française*, 87 (1990), p. 3-4.

DE LAMBERTERIE, *Introduction* = Ch. DE LAMBERTERIE, *Introduction à l'arménien classique*, dans *Lalies. Actes des sessions de linguistique et de littérature*, 10 (1992), p. 234-289.

DEREZA, *Lemmatization* = O. DEREZA, *Lemmatization for Ancient Languages: Rules or Neural Networks?*, dans *Artificial Intelligence and Natural Language. 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17-19, 2018, Proceedings*, ed. D. USTALOV – A. FILCHENKOV – L. PIVOVAROVA – J. ŽIŽKA, New York, NY – Dordrecht – Heidelberg – London, 2018, p. 35-47.

KINDT, *Principes* = B. KINDT, *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, dans *Byzantion*, 74 (2004), p. 213 272.

KINDT, *Processing Tools* = B. KINDT, *Processing Tools for Greek and Other Languages of the Christian Middle East*, in *Journal of Data Mining & Digital Humanities*, January 11, 2018, *Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*.

KINDT – HAELEWYCK – SCHMIDT – ATAS, *Concordance bilingue* = B. KINDT – J.-C. HAELEWYCK – A.B. SCHMIDT – N. ATAS, *La concordance bilingue grecque-syriaque des* Discours *de Grégoire de Nazianze*, dans *BABELAO*, 7 (2018), p. 51-80.

KINDT – PIRARD, *De Nazianze à Ninive* = B. KINDT – M. PIRARD, *De Nazianze à Ninive. La couverture lexicale du* Dictionnaire Automatique Grec, in V. SOMERS – P. YANNOPOULOS (eds.), *Philokappadox. In memoriam Justin Mossay* (*Orientalia Lovaniensia Analecta*, 25. *Bibliothèque de Byzantion*, 14), Louvain – Paris – Bristol, CT, 2016, p. 49-77.

KINDT – VIDAL-GORÈNE – DELLE DONNE, *Analyse automatique* = B. KINDT – Ch. VIDAL-GORÈNE – S. DELLE DONNE, *Analyse automatique du grec ancien par réseau de neurones. Évaluation sur le corpus* De Thessalonica Capta, dans *BABELAO*, 10-11 (2022), p. 525-550.

LAFONTAINE – COULIE, *Version* = G. LAFONTAINE – B. COULIE, *La version arménienne des discours de Grégoire de Nazianze. Tradition manuscrite et histoire du texte* (*Corpus Scriptorum Christianorum Orientalium*, 446; *Subsidia*, 67), Leuven, 1983.

MANJAVACAS – KÁDÁR – KESTEMONT, *Improving lemmatization* = E. MANJAVACAS – Á. KÁDÁR – M. KESTEMONT, *Improving lemmatization of non-standard languages with joint learning*, dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1 (Long and Short Papers), Minneapolis, MN, 2019, p. 1493-1503.

MEILLET, *Elementarbuch* = A. MEILLET, *Altarmenisches Elementarbuch* (*Indogermanische Bibliothek*, Erste Abteilung. *Sammlung Indogermanischer Lehr- und Handbücher*, I. Reihe. *Grammatiken*, 10), Heidelberg, 1913.

MEILLET, *Manuel élémentaire* = A. MEILLET, *Manuel élémentaire d'arménien classique,* traduction de G. KÉPÉKLIAN, Préface de Ch. DE LAMBERTERIE, Deuxième tirage revu et corrigé, Limoges, 2017.

SILBERZTEIN, *Dictionnaires* = M. SILBERZTEIN, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX* (*Informatique Linguistique*), Paris – Milan – Barcelone – Bonn, 1993.

TUERLINCKX, *La lemmatisation* = L. TUERLINCKX, *La lemmatisation de l'arabe non classique*, in *Le poids des mots. 7es Journées internationales d'Analyse statistique des Données Textuelles, 10-12 mars 2004, Louvain-la-Neuve*, ed. A. DISTER – C. FAIRON – G. PURNELLE, vol. II, Louvain-la-Neuve, 2004, p. 1069-1078.

VIDAL-GORÈNE – DECOURS-PEREZ, *Language resources* = Ch. VIDAL-GORÈNE – A. DECOURS-PEREZ, *Language resources for poorly endowed languages*: *The case study of Classical Armenian*, dans *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020, Marseille),* Paris, p. 3145–3152.

VIDAL-GORÈNE – KINDT, *From manuscript to tagged corpora* = Ch. VIDAL-GORÈNE – B. KINDT, *From manuscript to tagged corpora. An automated process for Ancient Armenian or other under resourced languages of the Christian East*, in *Armeniaca*, 1 (2022) (paper in submission).

VIDAL-GORÈNE – KINDT, *Lemma- and POS-tagging* = Ch. VIDAL-GORÈNE – B. KINDT, *Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old Georgian and*

Syriac, dans R. SPRUGNOLI – M. PASSAROTTI, *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020), Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, Paris, 2020, p. 22-27.

Université catholique de Louvain — Bernard COULIE
Institut orientaliste (CIOL) — Bastien KINDT
Place Blaise Pascal, 1 (Box L3.03.32) — Gabriel KÉPÉKLIAN
B-1348 Louvain-la-Neuve — Emmanuel VAN ELVERDINGHE
Belgique
bernard.coulie@uclouvain.be
bastien.kindt@uclouvain.be
gabriel.kepeklian@uclouvain.be
emmanuel.vanelverdinghe@uclouvain.be

Abstract — The GRE*g*ORI project provides scholars with lemmatized corpora of texts written in Greek and in the main languages of the Christian East. Attested word-forms are linked with lemma, POS-, and inflectional tags. This work makes it possible to produce lemmatized indexes and concordances, and to disseminate these corpora by using web-based interfaces. This paper gives an overview of the goals of the GRE*g*ORI project and lists the morphosyntactic and inflectional tags used for the analysis of the ancient Armenian language.