Modèles d'annotations morphologiques pour le traitement de données multivariées de l'arménien

Chahan Vidal-Gorène, Victoria Khurshudyan, Anaïd Donabédian

Modèles d'annotations morphologiques pour le traitement de données multivariées de l'arménien

Chahan Vidal-Gorène[1] Victoria Khurshudyan[2] Anaïd Donabédian[2]
(1) École Nationale des Chartes-PSL, 65 rue Richelieu, 75002 Paris, France
(2) SeDyL, UMR8202, INALCO, CNRS, IRD, 65 rue des Grands Moulins, 75013 Paris, France
chahan.vidal-gorene@chartes.psl.eu, victoria.khurshudyan@inalco.fr,
anaid.donabedian@inalco.fr

RÉSUMÉ

L'arménien est une langue comprenant de multiples variantes très inégales en termes de ressources disponibles en TAL. Nous avons entraîné un RNN pour réaliser l'annotation morphologique de différentes variantes de l'arménien, afin d'en comparer les résultats avec une approche par règles. Plusieurs tests ont permis d'évaluer la réutilisation d'un modèle non spécialisé de lemmatisation et de POS-tagging pour des variétés linguistiques sous représentées. Notre recherche s'est concentrée sur trois dialectes et a été étendue à l'arménien occidental, avec une précision moyenne de 94,00% en lemmatisation et 97,02% en POS-tagging, ainsi que sur une éventuelle réutilisation des modèles pour couvrir différentes autres variétés de l'arménien (jusqu'à 81% en POS-tagging). Nous montrons qu'une approche par RNN peut être une alternative valable à une approche par règles dans le cas d'une langue peu dotée et multivariées, en tenant compte de facteurs tels que la rapidité de traitement, la réutilisabilité pour différentes variétés d'une langue, et du gain qualitatif significatif en annotation morphologique.

ABSTRACT

Morphological Annotation Models for Armenian Multivariational Data Processing

Armenian is a language with significant variation and unevenly distributed NLP resources for different varieties. An attempt is made to process an RNN model for morphological annotation on the basis of different Armenian data and to compare the annotation results of RNN and rule-based models. Different tests were carried out to evaluate the reuse of an unspecialized model of lemmatization and POS-tagging for under-resourced language varieties. The research focused on three dialects and further extended to Western Armenian with a mean accuracy of 94,00% in lemmatization and 97,02% in POS-tagging, as well as a possible reusability of models to cover different other Armenian varieties (up to 81% in POS-tagging). It is argued that an RNN-based model can be a valid alternative to a rule-based one giving consideration to such factors as time-consumption, reusability for different varieties of a target language and significant qualitative results in morphological annotation results.

MOTS-CLÉS : variation linguistique, linguistique de corpus, arménien, RNN, règles-dictionnaire, étiquetage morphologique.

KEYWORDS: linguistic variation, corpus linguistics, Armenian, RNN, rule-based, morphological tagging.

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 71–81,

1 Introduction

Le renouvellement des ressources du traitement automatique des langues (TAL) au moyen de diverses ressources d'intelligence artificielle et réseaux de neurones (comme les réseaux de neurones récurrents, ci-après RNN) ainsi que le traitement de la variation linguistique présente un défi considérable en linguistique et en TAL pour les langues peu dotées. L'article explore et compare les différentes approches et méthodologies de lemmatisation et d'annotation linguistique pour un corpus multivariationnel, plus particulièrement pour un corpus combinant une variation diachronique et une variation synchronique [1]. L'annotation d'un corpus peut être réalisée via deux approches :

1. une approche classique dite par règles-dictionnaires, qui fait appel à un ensemble de règles pré-établies associées à un dictionnaire des formes annotées. Cette approche nécessite un temps de développement très important et les règles créées sont propres à un état de langue, sur lequel elle est très performante (Dereza, 2018) ;
2. une approche avec des réseaux de neurones, entraînés sur des corpus déjà annotés et qui peuvent réaliser des prédictions sur un nouveau corpus. Les RNN permettent en particulier de réaliser des annotations sur des tokens inconnus, qui sont prédominants dans le cas de corpus très différents, et de proposer des annotations contextuelles (Dereza, 2018). Il s'agit d'une approche gourmande en ressources et en données, mais qui, dans le cas de langues peu dotées, peut constituer une alternative intéressante et efficace à l'approche par règles-dictionnaires.

Jusqu'à présent, l'annotation linguistique (lemmatisation, POS-tagging, annotation morphologique et lexicale) des différentes variétés de l'arménien a principalement été réalisée avec des approches par règles-dictionnaires. Ces approches ont prouvé leur efficacité lorsque le système descriptif est suffisamment complet (dictionnaire), précis (règles) et le corpus traité homogène du point de vue de la variation linguistique (Khurshudyan et al., 2021) [pour l'arménien oriental], mais elles s'avèrent vite limitées et perturbées par différents facteurs identifiés (Vidal-Gorène and Kindt, 2020) [pour l'arménien classique]. Bien que souvent pertinentes, les approches par règles-dictionnaires souffrent de l'important temps de développement, et donc de ressources humaines, qu'elles nécessitent et de leur manque de polyvalence, en particulier pour des variations d'une même langue. Constituer des corpus annotés des différentes variations de l'arménien est en enjeu essentiel pour l'étude de cette langue, mais les annoter à la main ou créer un système règles-dictionnaire pour chacun d'entre eux est difficilement envisageable.

La présente recherche vise donc à explorer une alternative reposant sur des RNN pour annoter des variantes de l'arménien avec un même modèle. Cette approche a été favorisée pour sa flexibilité et son déploiement rapide, aussi bien linguistiquement que structurellement, sur des datasets variés. Les RNN permettent en particulier de réaliser des annotations sur des tokens inconnus, qui sont prédominants dans le cas de corpus multi-variantes, et de proposer des annotations contextuelles (Dereza, 2018). Des alternatives reposant sur des RNN ont déjà été testées pour l'arménien oriental (Arakelyan et al., 2018; Yavrumyan, 2019), et pour l'arménien classique (Vidal-Gorène and Kindt, 2020), et obtiennent des résultats très variables selon la spécialisation du registre du texte annoté, la taille de la base de données d'apprentissage et le pourcentage de tokens inconnus. Une tentative d'annotation morphologique de différentes ressources en arménien (issues de corpus morphologiquement annotés ou non) est ainsi

1. Les analyses de l'article reprennent partiellement les résultats d'une plus grande recherche visant à créer un corpus unifié diachronique et variationnel de la langue arménienne et plus particulièrement centrée sur le traitement des dialectes arméniens (Vidal-Gorène et al., 2020).

réalisée dans une démarche comparative avec les approches précédentes, afin d'évaluer le possible réemploi d'un modèle spécifique de lemmatisation et de POS-tagging sur des variantes linguistiques diverses et peu dotées, et les conditions d'une telle réutilisation.

2 La langue arménienne et état de l'art de ses ressources TAL

L'arménien est une langue indo-européenne à alignement nominatif-accusatif. Dans ses variantes modernes, la langue est morphologiquement dotée d'un système nominal essentiellement agglutinant et d'un système verbal plus fusionnel. Syntaxiquement, c'est une langue à tête finale, dont l'ordre des mots est flexible au niveau de la proposition (SVO / SOV).

La langue arménienne connaît d'importantes variations, que l'on peut schématiquement représenter par un axe horizontal diachronique allant de l'arménien classique (V^{e}-XII^{e} siècles) et l'arménien moyen (XI^{e}-XVII^{e} siècles) à l'arménien moderne (XVII^{e} siècle – aujourd'hui), et par un cercle décrivant le continuum synchronique arménien composé des deux standards de l'arménien moderne (oriental et occidental, standardisés au 19e siècle), de nombreux dialectes et variétés vernaculaires [pour plus d'informations sur la variation linguistique de l'arménien et son contexte géographique voir (Donabédian, 2018; Donabédian and Sitaridou, 2021)].

L'intercompréhension entre les variétés concernées peut varier d'une intelligibilité partielle à nulle. Ainsi, l'arménien classique est une langue flexionnelle avec une riche morphologie (presque totalement inintelligible pour le locuteur contemporain non-initié), l'arménien moyen est intermédiaire entre l'arménien classique et les variantes modernes, et se caractérise par une prolifération des paradigmes morphologiques, ainsi que l'abondance de lexique emprunté (presque totalement inintelligibles pour le locuteur contemporain), et enfin l'arménien occidental moderne et l'arménien oriental moderne et les trois dialectes choisis ont certains traits convergents, notamment une morphologie plus agglutinante, mais si l'intercompréhension entre les deux standards est relativement importante, a fortiori à l'écrit, les dialectes restent largement incompréhensibles pour les locuteurs des standards.

Bien que doté d'une tradition écrite multiséculaire, l'arménien manque significativement de ressources numériques. Plusieurs projets importants se consacrent au développement de ressources TAL pour des variétés arméniennes spécifiques participant d'une dynamique générale ascendante.

Les ressources existantes sont hétérogènes en termes d'accessibilité, de formatage et d'arrière-plan linguistique, lorsqu'elles existent pour une variante, elles ne la couvrent que très partiellement (à l'exception de l'EANC, qui vise à l'exhaustivité pour l'arménien oriental moderne), la majorité des variantes de l'héritage linguistique arménien sont totalement absentes, et les ressources de base du TAL sont encore plus rare.

Arménien classique : Les ressources les plus importantes de l'arménien classique comprennent la Bibliothèque numérique de littérature arménienne (Digital Library of Armenian Literature - Digilib) [2] avec une base de données textuelle simple, sans annotations linguistiques ; le corpus de la Bible (environ 630 000 tokens avec 60 000 tokens uniques et 12 000 lexèmes) avec annotation morphologique complète et alignement anglais réalisé par la fondation Arak29 [3] ; un corpus linguistique plus modeste (66 812 tokens avec 16 000 tokens uniques) spécialisé dans les textes hellénophiles classiques

2. www.digilib.am
3. www.arak29.am

arméniens (6ème-7ème siècles) établi par le projet GREgORI (UCLouvain) [4], le projet Calfa avec sa plateforme complète de dictionnaire de référence d'arménien classique en ligne (1,3 million de tokens, 190 000 tokens uniques) [5], et plusieurs autres bases de données d'arménien classique (projet TITUS [6], Leiden Armenian Lexical Textbase [7]).

Arménien moyen : Aucun corpus dédié à l'arménien moyen et à l'arménien occidental moderne n'est actuellement disponible.

Arménien moderne occidental : Digilib possède la plus importante base de données en texte brut de fiction et de textes historiques de l'arménien occidental moderne des XIX^{e} et XX^{e} siècles.

Arménien moderne oriental : La plus grande ressource pour l'arménien oriental est le Corpus national de l'arménien oriental (EANC) [8], un corpus exhaustif en libre accès avec environ 110 millions de tokens (du milieu du XIX^{e} siècle à aujourd'hui) et une annotation morphologique complète utilisant une approche fondée sur des règles. Le laboratoire YerevaNN et les UD [9] fournissent un échantillon annoté de l'arménien oriental moderne sous la forme d'une banque complète d'arbres de dépendance (53 000 tokens).

Dialectes arméniens : Le seul corpus dialectal [10] accessible en ligne a été conçu dans le cadre du projet de recherche EANC (environ 40 heures d'enregistrements et 250 000 tokens).

3 Datasets, expériences et méthodologies

Cinq principaux datasets [11] ont été constitués pour nos expérimentations : trois sur des variantes dialectales documentées et transcrites dans le cadre du projet de recherche EANC, et deux correspondant à chacun des standards de l'arménien moderne. Trois datasets mixtes ont été établis pour évaluer les gains potentiels en combinant les ressources. Ces données proviennent de la base de données de EANC.

1. **D-Ab** : Ce dataset est composé de transcriptions orales (15 heures, 16 informants) du dialecte d'Arcvaberd (Shamshadin, région de Tavush). Il comprend 120 258 wordforms (14 405 uniques) annotées manuellement. Il s'agit du dataset dialectal le plus important mais aussi le moins varié, avec seulement 4 120 lemmes uniques. **D-Ab** comporte de très nombreuses formes ambiguës, c'est-à-dire dont l'annotation définitive nécessite un arbitrage en fonction du contexte.
2. **D-Ga** : Ce dataset est composé de transcriptions orales (15 heures, 26 informants) du dialecte de Gusana (Maralik, région de Shirak). Il est composé de 100 352 wordforms (20 647 uniques) annotées manuellement. Avec un volume équivalent à **D-Ab**, il est beaucoup plus varié avec 9 087 lemmes uniques. En conséquence, nous trouvons beaucoup plus de tokens inconnus dans l'ensemble des tests qui lui est associé et **D-Ga** constitue donc un repère intéressant pour

4. www.gregoriproject.com
5. www.calfa.fr
6. www.titus.uni-frankfurt.de/indexe.htm
7. www.sd-editions.com/LALT/home.html
8. www.eanc.net
9. www.universaldependencies.org/treebanks/hy_armtdp/index.html
10. www.web-corpora.net/EANC_dialects/search
11. Pour décrire les résultats, nous définissons les acronymes suivants : **D-** pour dataset, associé à la langue concernée (p. ex. **D-MEA**). De même, les modèles créés sont nommés par : **m-** pour le modèle, associé à la langue concernée (p. ex. **m-MEA**).

l'évaluation des prédictions sur des tokens inconnus.

3. **D-Shn** : Ce dataset, composé de transcriptions orales (15 heures, 18 informants) du dialecte de Shenavan (Aparan, région de Aragatsotn), est le plus réduit des trois datasets. Il est composé de 89 632 wordforms (17 940 uniques) annotées manuellement. Il s'agit proportionnellement du dataset le plus varié, avec 7 568 lemmes uniques, et donc de nombreuses formes ambiguës et inconnues.
4. **D-MEA** : Il s'agit du dataset de référence de cet article, en arménien moderne oriental, sous-ensemble de EANC. Il est composé de 5 111 614 wordforms (201 710 uniques). Les phrases sont issues de sources hétérogènes, de la presse arménienne (2 037 629 wordforms), de fictions (1 453 894 wordforms) et de non-fictions (2 031 055 wordforms). **D-MEA** est représentatif de l'arménien moderne oriental.
5. **D-MWA** : Il s'agit d'un dataset expérimental en arménien moderne occidental, destiné à évaluer la pertinence de la réutilisation et reproductibilité des modèles. Il est composé de 3 531 wordforms (1 788 uniques).

Trois datasets mixtes ont été créés : **D-Ab+MEA**, composé de **D-Ab** et augmenté d'un tiers de son volume avec des données variées de **D-MEA**, ainsi que **D-Ga+MEA** et **D-Shn+MEA** augmentés selon le même procédé.

Enfin, trois autres datasets externes ont été considérés pour la création de modèles : les Universal Dependencies (**D-UD**) pour l'arménien oriental, et les données de GREgORI (**D-CA1**) et d'Arak29 (**D-CA2**) pour l'arménien classique.

Les annotations pour **D-Ab**, **D-GA** et **D-Shn** sont réalisées hors contexte (application d'une simple correspondance entre la liste des wordforms et le corpus), contrairement à **D-UD**, **D-CA1** et **D-CA2**. Par ailleurs, le paramètre de la graphie intervient dans la performance des annotations : **D-Ab**, **D-Ga**, **D-Shn**, **D-MEA**, **D-UD** utilisent une orthographe réformée, tandis que **D-MWA**, **D-CA1** et **D-CA2** l'orthographe classique. Des conversions orthographiques ont été réalisées pour permettre l'évaluation : si elles s'avèrent précises au niveau des mots, certaines constructions verbales n'existant pas dans toutes les variations, il en résulte parfois la création de bruit supplémentaire. De plus, **D-Ab**, **D-GA**, et **D-Shn** ont été transcrits en prenant en compte certaines spécificités phonologiques, ce qui ajoute une divergence graphique supplémentaire et affecte l'évaluation des différents modèles.

	D-MEA	**D-Ab**	**D-Ga**	**D-Shn**	**D-MWA**
Mots	5 111 614	120 258	100 352	89 632	3 531
Tokens uniques	201 710	14 405	20 647	17 940	1 788
Lemma uniques	-	4 120	9 087	7 568	1 311
Tokens ambigus	-	18 584	12 883	14 844	250
Tokens inconnus (test dataset)	13 145	364	1 968	1 810	1 080

TABLE 1 – Composition des principaux datasets

Trois types de réseaux ont été entraînés et évalués :

1. modèle RNN univariationnel pour une variété ciblée ;
2. modèle mixte (modèle 2/3 dialecte +1/3 arménien oriental moderne) ;
3. modèle RNN univariationnel pour une variété non-ciblée.

L'architecture RNN utilisée repose sur Pie (Manjavacas et al., 2019), qui offre une architecture très modulaire particulièrement adaptée au traitement des langues anciennes et variées, ce qui est majoritairement le cas ici et déjà éprouvée avec succès sur l'arménien classique (Vidal-Gorène and Kindt, 2020), dont nous reprenons l'essentiel du processus. Nous avons limité la capacité d'apprentissage de Pie — qui exploite pleinement le contexte d'une phrase pour améliorer les tâches de lemmatisation et de POS-tagging, en particulier en cas de token ambigu (Eger et al., 2016; Sprugnoli et al., 2020) — en raison de l'ambiguïté non levée des annotations dans **D-Ab**, **D-Ga** et **D-Shn**. Le RNN réalise donc par défaut une prédiction de toutes les analyses possibles pour un mot. La sélection de l'analyse la plus probable intervient seulement dans un second temps avec un modèle de langue.

Concernant le POS-tagging, le décodeur linéaire avait produit des résultats meilleurs sur l'arménien classique, mais nous avons néanmoins de nouveau comparé avec le décodeur CRF fourni par MarMoT et LEMMING (Mueller et al., 2013; Müller et al., 2015) et qui a fait ses preuves sur des datasets équivalents lors de la dernière Evalatin Evaluation Campaign (Sprugnoli et al., 2020; Stoeckel et al., 2020).

Enfin, pour que la comparaison soit complète, nous avons évalué la pertinence de l'architecture utilisée sur une langue standard (arménien oriental moderne - MEA), sur deux datasets (**D-MEA** et **D-UD**). Entraîné avec **D-UD** 2.3, COMBO (Rybak and Wróblewska, 2018) est efficace à 88,05% en lemmatisation et 85,07% en POS-tagging (Arakelyan et al., 2018; Yavrumyan, 2019). Entraînée avec **D-UD** 2.6, l'architecture décrite ici obtient 91,56% en lemmatisation (74,35% pour les tokens ambigus et 61,85% pour les tokens inconnus) et 92,54% en POS-tagging (87,81% ambigus et 83,56% inconnus).

Nous avons obtenu les résultats suivants en lemmatisation (voir annexes figure 2).

Modèles spécialisés : Pour la lemmatisation générale de tous les tokens des dialectes, les résultats varient entre 92,05% et 97,69%. Dans le cas des tokens inconnus, ceux-ci varient grandement, de 46,52% à 66,87%. Le modèle **m-Ab** (entraîné avec **D-Ab** qui contient le plus de tokens) s'avère être le plus performant dans la tâche générale, mais le manque de variété de tokens et de lemmes conduit à de mauvaises prédictions sur des tokens inconnus, là où **m-Ga** et **m-Shn** s'avèrent plus robustes. La matrice de confusion montre que **m-Ab** échoue majoritairement sur les formes verbales qui ne présentent pas de particularité phonétique (dans la transcription). **m-Ga** et **m-Shn** génèrent quant à eux beaucoup plus de formes fautives pour un même token, en plus d'être pénalisés par la grande variété de prononciations reproduites dans les corpus. **m-MEA** est quant à lui très robuste mais souffre de l'ambiguïté de ses données. Conséquence : le modèle s'avère plus performant quand on lui demande de générer toutes les analyses possibles plutôt que d'une seule (contrairement à **m-UD**). Il est néanmoins efficace à 94,34% sur **D-UD**.

Modèles mixtes : L'ajout de données de **D-MEA** à **D-Ab**, **D-Ga** et **D-Shn** en apprentissage entraîne un gain certain pour le dialecte d'Arcvaberd (+ 0,64% en accuracy, mais surtout +4,4% en précision et recall), y compris pour la prédiction sur des tokens inconnus qui passe de 46,52% à 51,10%. En revanche, cela pénalise les dialectes de Gusana et Shenavan (voir infra Modèles non-spécialisés pour une piste d'explication).

Modèles non-spécialisés : Un modèle strictement entraîné sur l'arménien oriental (**m-MEA** et **m-UD**) ne permet pour l'instant pas une lemmatisation des dialectes. Il en est de même pour des modèles entraînés sur l'arménien classique (**m-CA1** et **m-CA2**). Néanmoins, ces résultats sont à nuancer, puisque **D-Ab**, **D-Ga** et **D-Shn** sont transcrits très différemment et en orthographe réformée.

De nouvelles expérimentations doivent être réalisées après une uniformisation des annotations. Linguistiquement proches, les dialectes Arcvaberd et Gusana n'atteignent pas les 50% de bonne lemmatisation (respectivement 49,47% et 46,38%), ce qui reste néanmoins meilleur que l'annotation de **D-Ab** par **m-Shn** (42,90%). **m-Shn** annote correctement **D-Ga** à 58,45%, tandis que **m-Ga** annote **D-Shn** à 52,32%. On remarque un gain lors de l'utilisation de modèles mixtes.

Compte tenu des limites exposées dans la lemmatisation, les résultats en POS-tagging sont beaucoup plus réguliers (voir annexes figure 3). Il n'a pas été possible de réaliser toutes les évaluations de POS-tagging en raison de la trop grande variation d'annotation des corpus (étiquettes utilisées et caractérisation). Ce dernier point constitue un enjeu important pour des expérimentations ultérieures.

Modèles spécialisés : Nous obtenons de très bons modèles de POS-tagging (> 95%) pour chacun des dialectes, y compris dans le cas de tokens inconnus. **m-Ab** reste peu robuste face à la diversité. Plus de deux tiers des erreurs sont localisés sur une confusion entre nom et adjectif, qui peut s'expliquer par l'absence de contexte.

Modèles mixtes : En POS-tagging, l'ajout de MEA aux dialectes apporte un vrai gain dans l'annotation, en particulier pour les tokens inconnus.

Modèles non spécialisés : La réutilisation des modèles d'un dialecte à un autre apparaît clairement possible ici, en particulier pour les dialectes Gusana et Shenavan, qui n'appartient pourtant pas à la même branche linguistique, avec une couverture de 82,22% de **D-Ga** par **m-Shn+MEA**. Par ailleurs, même seulement efficace à 66,27% (**m-Ga** sur **D-Ab**), cela peut apporter une bonne base pour l'annotation plus rapide de corpus dialectaux.

Le modèle **m-MEA** propose une lemmatisation correcte à 88,79% et un POS-tagging correct à 87,33% d'un corpus en arménien occidental moderne. Le parser de EANC obtient respectivement 74,09% et 68,57% sur ce même corpus.

4 Conclusion

Nous avons réalisé différentes expérimentations qui illustrent pour la première fois l'annotation automatique de variantes dialectales de l'arménien, et la réutilisation possible de modèles non-spécialisés pour la constitution rapide de corpora.

La précision moyenne d'annotation des RNN est de 94% en lemmatisation et 97,02% en POS-tagging. Ils ont notamment montré une grande polyvalence pour traiter différentes variétés linguistiques autres que leur base d'apprentissage, au contraire des systèmes traditionnels. Les tests réalisés mettent notamment en évidence une reproductibilité des modèles de l'ordre de 74%. Mais il a toutefois fallu les entraîner avec une base d'apprentissage préalablement annotée contenant plus de 5 millions de tokens. Entraînés sur des corpus plus modestes (60 000 tokens), les RNN proposent des résultats parfois en-deça des systèmes règles-dictionnaires (précision moyenne de 92%), bien que restant meilleurs sur les tokens inconnus (Vidal-Gorène and Kindt, 2020). En particulier, la comparaison d'un modèle RNN entraîné sur de l'arménien oriental et d'un modèle règles-dictionnaire en arménien oriental appliqués à de l'arménien occidental moderne montre un gain significatif de 19% en annotation. Le modèle RNN couvre en effet 88,79% d'un petit ensemble de données hétérogènes en arménien occidental, modèle qui pourrait donc être un point de référence pour l'annotation massive de corpus dans ce standard linguistique.

Nous soutenons ainsi qu'un modèle reposant sur des RNN peut être une alternative probante à une approche par règles, compte tenu de facteurs comme le temps de traitement, la réutilisation d'un même modèle sur différentes variétés d'une langue donnée, et les gains en annotation morphologique pour des langues peu dotées. Cela est d'autant plus pertinent dans le cadre d'une langue peu dotée, car il permet de construire rapidement une base d'apprentissage suffisante pour établir un RNN polyvalent. Les différentes études conduites, en complément de la mise à disposition des architectures neuronales ou hybrides construites, servent de point de repère pour l'annotation de langues peu dotées à graphies non latines. Une nouvelle approche, hybride, devrait pouvoir concilier rapidité d'application et couverture maximale de l'annotation. La présente recherche montre une possibilité de réutiliser des modèles pour couvrir d'autres variétés linguistiques, même partiellement.

Des recherches futures permettront d'étendre le corpus multivariationnel afin d'inclure autant de variantes linguistiques arméniennes que possible (des corpus visant à l'exhaustivité pour l'arménien classique, l'arménien moyen et l'arménien occidental moderne, ainsi que des dialectes arméniens et des variantes vernaculaires) et de tester et évaluer les modèles existants sur les nouvelles données.

Jusqu'à présent, la distance variationnelle entre variantes linguistiques a été calculée sur la base d'un faisceau de traits linguistiques. La question de la distance linguistique entre deux variétés est particulièrement pertinente dans la classification généalogique des langues du monde et dans le cadre de la classification des dialectes. Dans la tradition dialectologique arménienne, on a eu recours pour la classification des dialectes à divers critères linguistiques et extra-linguistiques (géographique, morphologique, phonologique, multiparamétrique, etc.). Appliquer le modèle d'annotation morphologique RNN d'une variété particulière sur d'autres variétés typologiquement proches (dychroniquement ou synchroniquement) et évaluer la distance variationnelle entre deux variétés sur la base de différents paramètres formels permettrait de formuler des hypothèses nouvelles et de revoir les classifications existantes.

Les variétés linguistiques ont généralement une vitalité fragile en l'absence de prestige social et / ou de renouvellement des locuteurs natifs. C'est en particulier le cas pour les dialectes arméniens qui souffrent d'un déficit de reconnaissance par rapport à la langue standard avec laquelle ils coexistent. Par conséquent, la documentation et l'annotation de ces variétés linguistiques sont d'un enjeu majeur en TAL mais aussi et surtout dans des perspectives linguistiques, anthropologiques et sociales.

Références

Arakelyan, G., Hambardzumyan, K., and Khachatrian, H. (2018). Towards JointUD : Part-of-speech Tagging and Lemmatization using Recurrent Neural Networks. In *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 180–186, Brussels, Belgium. Association for Computational Linguistics.

Dereza, O. (2018). *Lemmatization for Ancient Languages : Rules or Neural Networks ?*, pages 35–47. Springer International Publishing, Cham.

Donabédian, A. (2018). Middle East and Beyond - Western Armenian at the crossroads : A sociolinguistic and typological sketch. In Bulut, C., editor, *Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery*, volume 111 of *Turcologica*, pages 89–148. Harrazowitz Verlag, Wiesbaden, Allemagne.

Donabédian, A. and Sitaridou, I. (2021). Anatolia. In Adamou, E. and Matras, Y., editors, *The Routledge Handbook of Language Contact*, pages 404–433. Routledge, London, England.

Eger, S., Gleim, R., and Mehler, A. (2016). Lemmatization and Morphological Tagging in German and Latin : A comparison and a survey of the state-of-the-art. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1507–1513, Portorož, Slovenia. European Language Resources Association (ELRA).

Khurshudyan, V., Arkhangelskiy, T., Daniel, M., Levonian, D., Plungian, V., Polyakov, A., and Rubakov, S. (2021). Introduction to Eastern Armenian National Corpus : www.eanc.net. *Études arméniennes contemporaines*. submitted.

Manjavacas, E., Ákos, K., and Mike, K. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. In *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)*, pages 1493–1503, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Rybak, P. and Wróblewska, A. (2018). Semi-Supervised Neural System for Tagging, Parsing and Lematization. In *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium. Association for Computational Linguistics.

Sprugnoli, R., Passarotti, M., Cecchini, F. M., and Pellegrini, M. (2020). Overview of the EvaLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Stoeckel, M., Henlein, A., Hemati, W., and Mehler, A. (2020). Voting for POS tagging of Latin texts : Using the flair of FLAIR to better Ensemble Classifiers by Example of Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135, Marseille, France. European Language Resources Association (ELRA).

Vidal-Gorène, C., Khurshudyan, V., and Donabédian, A. (2020). Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 90–101, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Vidal-Gorène, C. and Kindt, B. (2020). Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old georgian, and Syriac. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 22–27, Marseille, France. European Language Resources Association (ELRA).

Yavrumyan, M. (2019). Tokenization and Word Segmentation in the UD ARMENIAN-ArmTDP Treebank. *Banber Erewani hamalsarani*, pages 52–65.

5 Annexes

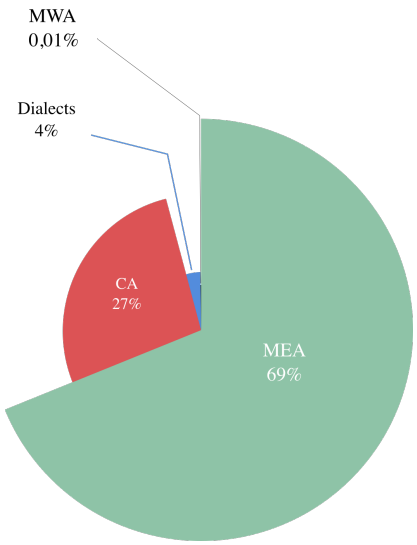

FIGURE 1 – Proportion par langue des ressources annotées et en open access, utilisées dans le cadre de cette étude.

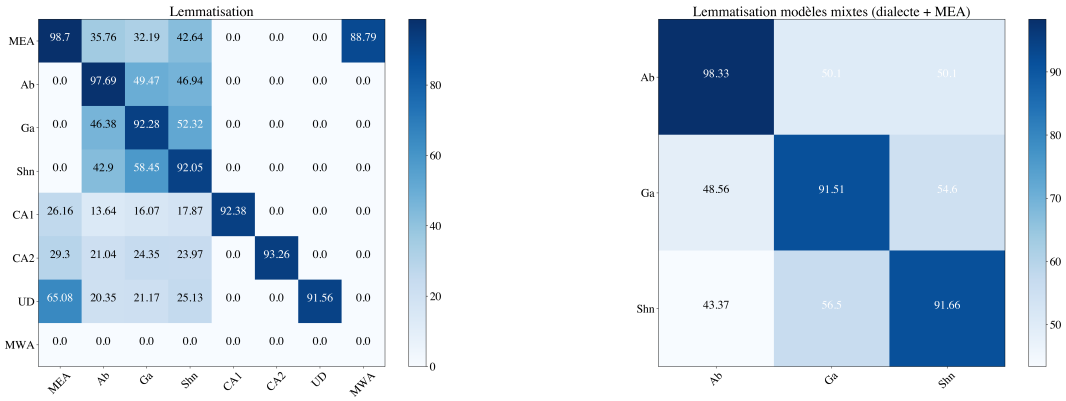

FIGURE 2 – Évaluation des modèles simples et mixtes de lemmatisation sur tous les tokens (accuracy) répartition des résultats. La valeur 0 indique que le modèle n'a pu être appliqué sur le dataset concerné, en raison des limites évoquées dans l'article.

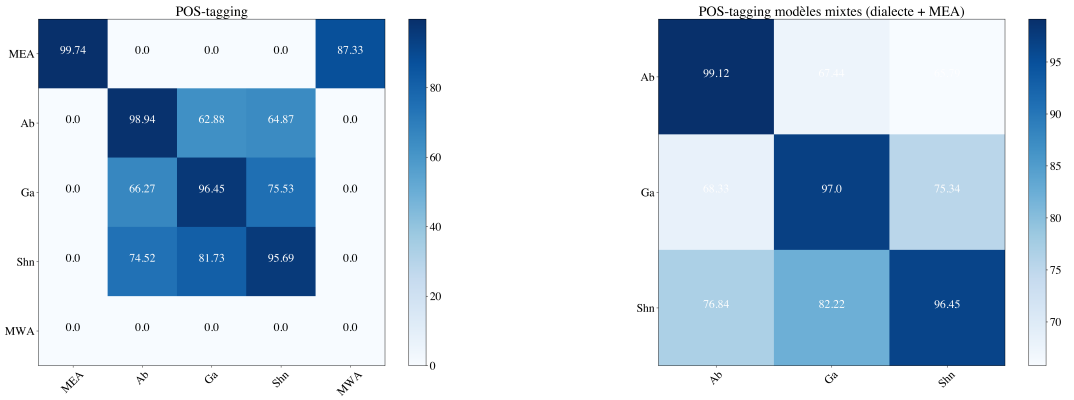

FIGURE 3 – Évaluation des modèles simples et mixtes de POS-tagging sur tous les tokens (accuracy) et répartition des résultats. La valeur 0 indique que le modèle n'a pu être appliqué sur le dataset concerné, en raison des limites évoquées dans l'article.