Transfert inter-dialectal et généralisation zéro-shot en arménien

Chahan Vidal-Gorène[1,2] Nadi Tomeh[1] Victoria Khurshudyan[3]
(1) LIPN, CNRS UMR 7030, France
(2) École nationale des chartes, PSL University, Centre Jean Mabillon, France
(3) SeDyL, UMR8202, INALCO, CNRS, IRD, France
chahan.vidal-gorene@chartes.psl.eu

RÉSUMÉ

Cet article évalue trois tâches de TAL (lemmatisation, étiquetage morphosyntaxique, analyse morphologique) sur quatre variétés de l'arménien : arménien classique, oriental moderne, occidental moderne et le dialecte de Getashen. Trois familles de modèles sont comparées : réseaux récurrents (RNN), transformer multilingue (mDeBERTa) et modèle génératif (GPT-4-Turbo-2024-04-09). Les RNNs excellent en supervision complète, mDeBERTa modélise efficacement la morphologie, tandis que GPT-4-Turbo surpasse les autres en transfert zéro- et few-shot vers Getashen. Ces résultats soulignent l'efficacité de l'apprentissage *in-context* pour le TAL des langues peu dotées et des variétés dialectales.

ABSTRACT

Armenian NLP across Varieties : Cross-Dialectal Transfer and Zero-Shot Learning

We evaluate lemmatization, POS-tagging, and morphological analysis across four Armenian varieties : Classical Armenian (CA), Modern Eastern Armenian (MEA), Modern Western Armenian (MWA), and the under-documented Getashen dialect (G). Three model families are compared : RNNs, mDeBERTa, and GPT-4-Turbo-2024-04-09. RNNs perform best in supervised settings, mDeBERTa captures morphology across standards, while GPT-4-Turbo shows superior zero- and few-shot transfer to Getashen. These results highlight in-context learning as a powerful strategy for cross-dialectal and low-resource NLP.

MOTS-CLÉS : arménien, transfert inter-variétés, zéro-shot, TAL, langues peu dotées.

KEYWORDS: Armenian, cross-dialect transfer, zero-shot, NLP, low-resource languages.

1 Introduction

Le traitement automatique du langage (TAL) appliqué à l'arménien se heurte à deux obstacles majeurs : la faible disponibilité de ressources et une forte variation dialectale. Si l'arménien oriental (MEA) dispose désormais de corpus annotés et de jeux d'évaluation (Khurshudyan *et al.*, 2022a,b), l'arménien occidental (MWA) reste peu exploré et doté, l'arménien classique

(CA) repose essentiellement sur des sources historiques (Vidal-Gorène *et al.*, 2020; Vidal-Gorène & Kindt, 2020), et des variétés orales comme le dialecte de Getashen (G) demeurent très peu documentées (Khurshudyan & Shagoyan, 2016). La constitution de corpus annotés pour chacune de ces variétés s'avère coûteuse et souvent irréalisable, compte tenu de la rareté des locuteurs et/ou des données de terrain disponibles (plans de numérisation, etc.). Il faut donc privilégier des approches qui réutilisent les corpus annotés existants pour permettre le transfert inter-variétés et l'apprentissage en faible supervision.

Dans le cadre du projet ANR DALiH[1], nous examinons dans quelle mesure le transfert inter-dialectal et les LLM en situation de *zero-shot* ou *few-shot* peuvent compenser le manque de données annotées d'une variété cible. Nous comparons pour cela trois familles de modèles : des réseaux récurrents (RNN), un transformeur et un modèle génératif[2].

2 Aperçu linguistique de l'arménien

L'arménien est une langue indo-européenne constituant une branche à part entière de cette famille de langues. Elle présente une variation importante, tant historique que géographique. Traditionnellement, son histoire est divisée en trois stades : l'arménien classique (V^{e}–X^{e} siècles), l'arménien moyen (XI^{e}–XVI^{e} siècles) et l'arménien moderne (à partir du XVII^{e} siècle), ce dernier comprenant deux normes écrites — l'arménien oriental et l'arménien occidental — ainsi qu'un ensemble de dialectes régionaux. Toutes les variétés écrites utilisent l'alphabet arménien. La présente étude porte sur l'annotation morphologique et morphosyntaxique de quatre variétés : l'arménien classique, l'arménien occidental, l'arménien oriental et le dialecte de Getashen.

L'arménien classique est une langue à tête initiale, fortement flexionnelle, à alignement nominatif–accusatif et à ordre des mots relativement libre. Son corpus couvre des textes religieux, historiographiques, scientifiques ou médicaux, même si son usage contemporain reste essentiellement liturgique. À l'inverse, les variétés modernes (MEA, MWA) et le dialecte de Getashen présentent une structure à tête finale, conservent un alignement nominatif–accusatif, et tendent vers une morphologie plus agglutinante, avec un usage plus fréquent des constructions périphrastiques et une syntaxe souple. Les deux normes modernes ont été standardisées au milieu du XIX^{e} siècle, donnant lieu à une production écrite importante : le MWA est majoritairement utilisé par la diaspora, tandis que le MEA domine en Arménie, en Iran et dans les communautés issues de l'ex-URSS.

Le dialecte de Getashen appartient à la branche dite *-um* des parlers du Karabagh (Martirosyan, 2019; Davtyan, 1966) et est essentiellement oral. Les données utilisées ici proviennent d'enquêtes de terrain menées en Arménie entre 2014 et 2015, dans le cadre du projet *Mi-*

1. https ://dalih.fr – Digitizing Armenian Linguistic Heritage : Armenian Multivariational Corpus and Data Processing, ANR-21-CE38-0006
2. Version actualisée et traduite en français de Vidal-Gorène *et al.* (2024).

gration and Complex Identities in the Republic of Armenia (Khurshudyan & Shagoyan, 2016).

3 Données utilisées

Cette étude mobilise quatre ensembles de données. Trois d'entre eux proviennent du projet **Universal Dependencies (UD)** (De Marneffe *et al.*, 2021) : CA [3], MEA [4] et MWA [5]. Le quatrième corpus, correspondant au dialecte de **Getashen (G)**, est un échantillon extrait du projet *Migration and Complex Identities in the Republic of Armenia* (Khurshudyan & Shagoyan, 2016).

Les corpus UD offrent une annotation morphologique et syntaxique détaillée conforme aux lignes directrices du projet. En revanche, le corpus de Getashen se compose de transcriptions d'entretiens oraux, dont seule une partie a été annotée manuellement pour les besoins de cette étude.

Variété	Tokens	Phrases	Source / Domaine	Annotation
CA	82 557	4 146	Évangiles (PROIEL → UD)	Lemme, POS, Morph, Syntaxe
MEA	∼52 950	2 500	Blogs, fiction, juridique, presse	Lemme, POS, Morph, Syntaxe
MWA	∼124 230	6 656	Genres mixtes (manuel + auto)	Lemme, POS, Morph, Syntaxe
G	∼1 200	100	Données orales (normes EANC)	Lemme, POS (manuel)

TABLE 1 – Jeux de données utilisés pour chaque variété de l'arménien.

4 Méthodologie

Nous comparons trois familles de modèles : (i) des **réseaux de neurones récurrents (RNN)**, implémentés avec l'architecture *PIE* (Manjavacas *et al.*, 2019), qui constituent une référence robuste pour les tâches d'analyse linguistique en contexte de faible ressource (Vidal-Gorène & Kindt, 2020) ; (ii) un **transformeur multilingue**, *mDeBERTa*, offrant un potentiel de transfert interlingue et inter-variétés (Kondratyuk, 2019) ; (iii) un **LLM génératif** (*GPT-4-Turbo-2024-04-09*), évalué en configurations *zero-* et *few-shot* (Bansal & Sharma, 2023; Kholodna *et al.*, 2024).

Trois tâches sont considérées : la lemmatisation, l'étiquetage morphosyntaxique (POS-tagging) et la prédiction de la morphologie. Nous rapportons les scores de macro-F1 pour l'apprentissage supervisé sur les corpus UD (CA, MEA, MWA) et pour le transfert vers le dialecte de Getashen. Les ensembles UD complets sont utilisés pour l'entraînement et l'évaluation *in-domain*, tandis que le sous-ensemble Getashen (100 phrases) sert de jeu de test *out-of-domain*.

3. https ://universaldependencies.org/treebanks/xcl_caval/
4. https ://universaldependencies.org/treebanks/hy_armtdp/
5. https ://universaldependencies.org/treebanks/hyw_armtdp/

Dans les expériences *few-shot* avec GPT-4, de petits échantillons de 10, 50, 100 et 500 tokens par variété ont été fournis sous forme d'exemples *in-context*. Cette approche permet d'évaluer la capacité du modèle à généraliser sans ajustement de paramètres, uniquement à partir d'un nombre limité d'exemples directement apportés dans le prompt.

5 Résultats

En apprentissage supervisé, les RNN atteignent des performances quasi parfaites en étiquetage morphosyntaxique (≈0.98 F1) et des précisions élevées en lemmatisation (0.66–0.91) sur les variétés CA, MEA et MWA. Le modèle *mDeBERTa* obtient des résultats légèrement inférieurs (0.88–0.91 en POS), mais reste compétitif pour la modélisation morphologique (0.77–0.88). Ses performances plus faibles et instables sur le corpus CA reflètent une exposition limitée aux données historiques lors du pré-entraînement multilingue.

En transfert vers le dialecte non vu de Getashen, les deux modèles montrent une forte chute de performance : les RNN n'atteignent que 0.11–0.13 en F1 pour le POS et jusqu'à 0.37 pour la lemmatisation, tandis que *mDeBERTa* obtient entre 0.34–0.62 et 0.08–0.26 respectivement, confirmant la difficulté du transfert inter-variétés dans ce cadre supervisé, sur un corpus syntaxiquement très différent.

À l'inverse, *GPT-4-Turbo* – sans aucun *fine-tuning* – atteint 0.86 F1 en zéro-shot sur l'étiquetage POS du dialecte Getashen, et 0.83 en lemmatisation, avec une amélioration au-delà de 0.90 en configuration *few-shot in-context*.

Modèle	***POS tagging* (F1)**		**Lemmatisation (F1)**		**Morphologie (CA/MEA/MWA)**
	CA/MEA/MWA	Getashen (transfert)	CA/MEA/MWA	Getashen (transfert)	
RNN (*fine-tuned*)	0.98	0.11–0.13	0.66–0.91	0.14–0.37	0.70–0.88
mDeBERTa (*fine-tuned*)	0.88–0.91	0.34–0.62	0.36–0.70	0.08–0.26	0.77–0.88
GPT-4-Turbo (*zero-shot*)	0.86–0.91	0.86	0.62–0.83	0.83	0.71–0.84
GPT-4-Turbo (*few-shot*)	0.90+	0.89	0.74–0.83	0.90+	0.75–0.88

TABLE 2 – Scores macro-F1 pour l'étiquetage morphosyntaxique, la lemmatisation et l'analyse morphologique. Les modèles RNN et mDeBERTa sont *fine-tunés* sur les corpus UD (CA/MEA/MWA) et évalués en domaine et en transfert vers Getashen. GPT-4-Turbo est évalué en configurations *zero-shot* et *few-shot in-context* sans aucun ajustement de paramètres. L'analyse morphologique est mesurée uniquement pour les trois variétés UD. *Voir Vidal-Gorène et al. (2024) pour les résultats détaillés et les conditions expérimentales.*

6 Discussion : mises à jour spécifiques à l'arménien

Les résultats mettent en évidence la complémentarité des trois familles de modèles. Les *RNN* offrent des performances solides et stables en supervision complète, tandis que *mDeBERTa* modélise efficacement la morphologie et se transfère bien entre les deux normes modernes

(MEA–MWA, voir Vidal-Gorène *et al.* (2024)). Ces deux architectures se dégradent toutefois nettement sur des variétés non vues, comme le dialecte de Getashen, soulignant les limites des approches supervisées, même entre variantes d'une même langue.

À l'inverse, *GPT-4-Turbo* montre une forte capacité d'adaptation inter-variétés, atteignant les meilleurs scores de transfert sans mise à jour de paramètres et progressant encore avec un faible nombre d'exemples contextuels. L'apprentissage *in-context* en *few-shot* permet ainsi d'égaler, voire de dépasser, les performances supervisées en contexte de faible ressource, confirmant les résultats récents sur le *prompting* et l'adaptation interlingue (Bansal & Sharma, 2023; Kholodna *et al.*, 2024; Vidal-Gorène *et al.*, 2025).

Pour le TAL de l'arménien, trois pistes se dégagent : (1) les jeux de référence du MEA (Khurshudyan *et al.*, 2022b; Waldenfels von R. & Dobrushina, 2014) pourraient bénéficier de stratégies de *prompt design* adaptées aux jeux d'étiquettes ; (2) le MWA et les dialectes comme Getashen nécessitent davantage de données de terrain et d'annotation manuelle (Khurshudyan & Shagoyan, 2016; Arkhangelskiy & Georgieva, 2018) ; (3) de nouvelles ressources et benchmarks partagés (Riemenschneider & Krahn, 2024) rendent désormais possible l'évaluation directe de ces méthodes et la comparaison des résultats entre corpus.

Les approches hybrides — combinant des modules RNN ou transformeurs pour les tâches de base avec un LLM génératif tel que *GPT-4-Turbo* pour l'inférence adaptative — apparaissent particulièrement prometteuses, tant pour la recherche linguistique que pour les applications en humanités numériques et linguistiques (Kindt & Van Elverdinghe, 2022; Sahala, 2024).

7 Conclusion

Les RNN restent la référence pour l'étiquetage morphosyntaxique supervisé, et *mDeBERTa* offre une modélisation morphologique robuste entre standards. Aucun des deux, cependant, ne généralise efficacement à des dialectes non vus. Grâce au *prompting* en *zero-* et *few-shot*, *GPT-4-Turbo* réalise un transfert inter-dialectal performant, soulignant l'importance des approches hybrides et du développement de corpus variés comme prochaines étapes pour le TAL en contexte de faible ressource. Les pistes explorées ici pour l'arménien pourraient également s'appliquer à d'autres langues sous-dotées et plurivariationnelles.

Disponibilité des modèles

L'ensemble des modèles *PIE* utilisés dans cette étude est disponible sur Zenodo. Ils ont été publiés dans les actes de la conférence *EMNLP 2024 – NLP for Digital Humanities* (Vidal-Gorène, Tomeh & Khurshudyan, 2024) et couvrent respectivement trois variétés de l'arménien :

— **arménien classique** — *Pie Model for Lemmatization, POS Tagging, and Morphological Analysis of Classical Armenian*. https://doi.org/10.5281/zenodo.14056139
— **arménien oriental** — *Pie Model for Lemmatization, POS Tagging, and Morphological Analysis of Eastern Armenian*. https://doi.org/10.5281/zenodo.14059437
— **arménien occidental** — *Pie Model for Lemmatization, POS Tagging, and Morphological Analysis of Western Armenian*. https://doi.org/10.5281/zenodo.14060082

Références

ARKHANGELSKIY T. & GEORGIEVA E. (2018). Sound-aligned corpus of udmurt dialectal texts. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, p. 26–38 : Association for Computational Linguistics.

BANSAL P. & SHARMA A. (2023). Large language models as annotators : Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv :2306.15766*.

DAVTYAN K. (1966). *Leŕnayin Ġarabaġi barbaŕayin k'artez [= The dialectal map of Nagorno-Karabakh]*. Yerevan.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal dependencies. *Computational linguistics*, **47**(2), 255–308.

KHOLODNA N., JULKA S., KHODADADI M., GUMUS M. N. & GRANITZER M. (2024). Llms in the loop : Leveraging large language model annotations for active learning in low-resource languages. *arXiv preprint arXiv :2404.02261*.

KHURSHUDYAN V., ARKHANGELSKIY T., DANIEL M., PLUNGIAN V., LEVONIAN D., POLYAKOV A. & RUBAKOV S. (2022a). Eastern Armenian national corpus : State of the art and perspectives. In V. KHURSHUDYAN, N. TOMEH, D. NOUVEL, A. DONABEDIAN & C. VIDAL-GORENE, Éds., *Proceedings of the Workshop on Processing Language Variation : Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, p. 28–37, Marseille, France : European Language Resources Association.

KHURSHUDYAN V., ARKHANGELSKIY T., DANIEL M., PLUNGIAN V., LEVONIAN D., POLYAKOV A. & RUBAKOV S. (2022b). Eastern armenian national corpus : State of the art and perspectives. In *Proceedings of the Workshop on Processing Language Variation : Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, p. 28–37.

KHURSHUDYAN V. & SHAGOYAN G. (2016). Obscured borders of migrants' 'locality' : Language and identity shift of armenian refugees from azerbaijan : Case study of getashen village. In *Language Indexicality and Belonging Conference*.

KINDT B. & VAN ELVERDINGHE E. (2022). Describing language variation in the colophons of armenian manuscripts. In *Proceedings of the Workshop on Processing Language Variation : Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, p. 20–27.

KONDRATYUK D. (2019). Cross-lingual lemmatization and morphology tagging with two-stage multilingual bert fine-tuning. In *Proceedings of the 16th workshop on computational research in phonetics, phonology, and morphology*, p. 12–18.

MANJAVACAS E., KÁDÁR Á. & KESTEMONT M. (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)*, p. 1493–1503 : Association for Computational Linguistics.

MARTIROSYAN H. (2019). *2.2. The Armenian dialects*, In G. HAIG & G. KHAN, Éds., *The Languages and Linguistics of Western Asia*, p. 46–105. De Gruyter Mouton : Berlin, Boston. DOI : doi :10.1515/9783110421682-003.

RIEMENSCHNEIDER F. & KRAHN K. (2024). Heidelberg-boston@ sigtyp 2024 shared task : Enhancing low-resource language analysis with character-aware hierarchical transformers. *arXiv preprint arXiv :2405.20145*.

SAHALA A. (2024). Neural lemmatization and pos-tagging models for coptic, demotic and earlier egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, p. 87–97.

VIDAL-GORÈNE C., CAFIERO F. & KINDT B. (2025). Under-resourced studies of under-resourced languages : lemmatization and POS-tagging with LLM annotators for historical Armenian, Georgian, Greek and Syriac. working paper or preprint.

VIDAL-GORÈNE C., KHURSHUDYAN V. & DONABÉDIAN-DEMOPOULOS A. (2020). Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing. In M. ZAMPIERI, P. NAKOV, N. LJUBEŠIĆ, J. TIEDEMANN & Y. SCHERRER, Éds., *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 90–101, Barcelona, Spain (Online) : International Committee on Computational Linguistics (ICCL).

VIDAL-GORÈNE C. & 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*, p. 22–27.

VIDAL-GORÈNE C., TOMEH N. & KHURSHUDYAN V. (2024). Cross-dialectal transfer and zero-shot learning for armenian varieties : A comparative analysis of rnns, transformers and llms. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, p. 438–449.

WALDENFELS VON R. D. M. & DOBRUSHINA N. (2014). Why standard orthography ? building the ustya river basin corpus, an online corpus of a russian dialect. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, volume 13, p. 720–728.