Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of RNNs, Transformers and LLMs

Chahan Vidal-Gorène, Nadi Tomeh, Victoria Khurshudyan

Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of RNNs, Transformers and LLMs

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
Correspondence: chahan.vidal-gorene@chartes.psl.eu

Abstract

This paper evaluates lemmatization, POS-tagging, and morphological analysis for four Armenian varieties: Classical Armenian, Modern Eastern Armenian, Modern Western Armenian, and the under-documented Getashen dialect. It compares traditional RNN models, multilingual models like mDeBERTa, and large language models (ChatGPT) using supervised, transfer learning, and zero/few-shot learning approaches. The study finds that RNN models are particularly strong in POS-tagging, while large language models demonstrate high adaptability, especially in handling previously unseen dialect variations. The research highlights the value of cross-variational and in-context learning for enhancing NLP performance in low-resource languages, offering crucial insights into model transferability and supporting the preservation of endangered dialects.

1 Introduction

This research is part of the DALiH project[1]. The goal of the project is to create a robust digital platform for the exploration of both historical and contemporary forms of the Armenian language. The project intends to offer freely accessible and open-source resources, which will include grammatically annotated corpora and a variety of NLP tools such as annotation models, datasets, ASR models, among others.

This study focuses on four varieties of Armenian: Classical Armenian (CA), Modern Eastern Armenian (MEA), Modern Western Armenian (MWA), and the Getashen dialect (G). While CA, MEA, and MWA have annotated corpora and models included in the Universal Dependencies (UD) project, the Getashen dialect, which is an underdocumented variety[2]

Linguistic resources like annotated corpora and NLP models for tasks such as lemmatization, POS-tagging, and morphological analysis are essential for both linguists and digital humanities scholars. These tools support tasks like corpus pre-annotation and the study of historical texts, oral traditions, and regional literature. In this study, we aim to develop models for these tasks across the four varieties, with a particular focus on the under-resourced Getashen dialect.

Our contributions are threefold:

1. Comparative Evaluation of Models: We conduct a comprehensive comparative study of traditional RNN models, state-of-the-art multilingual language models (e.g., mDeBERTa), and large models (e.g., ChatGPT) in zero and few-shot setups across the three UD-supported dialects.

2. Pre-annotation of the Getashen Dialect: We evaluate the feasibility of using existing NLP models to pre-annotate the Getashen dialect, assessing the adaptability and transferability of models trained on other dialects.

3. Variational divergences / convergences: We explore linguistic similarities between the dialects and how they affect model transferability, providing insights into cross-dialectal model adaptation.

2 Linguistic Background

Armenian is an Indo-European language that constitutes a distinct branch marked by both diachronic

[1]The DALiH project is funded by French National Research Agency ANR-21-CE38-0006.: Digitizing Armenian Linguistic Heritage: Armenian Multivariational Corpus and Data Processing

[2]In our study, we employ two terms to characterize the status of linguistic data and resources: an *"underdocumented language,"* which denotes a language or variety that lacks formal linguistic records, and an *"under-resourced language,"* which signifies a language or variety that lacks digital resources.

and synchronic variation. The historical evolution of Armenian comprises: a) Classical Armenian (5th-10th centuries A.D.), b) Middle Armenian (11th-16th centuries), and c) Modern Armenian (17th century to the present). Synchronically, Modern Armenian includes Modern Eastern Armenian (MEA), Modern Western Armenian (MWA) and numerous Armenian dialects. All the written forms of the Armenian language utilize the unique Armenian alphabet.

While the linguistic divergences in historical variation are considerable, they vary in degree among the two modern standards and dialectal varieties, depending on the areal and genetic distance of each within the Armenian linguistic continuum (for more details, see Sayeed and Vaux (2017); for linguistic variation, see Donabedian-Demopoulos (2018) and Khurshudyan and Donabédian (2021)).

This study explores the morphological and morphosyntactic annotation of the following Armenian linguistic varieties: Classical Armenian, Modern Western Armenian, Modern Eastern Armenian and the Getashen dialect. Classical Armenian (CA) is typologically a right-branching language with complex inflectional morphology and syntactic constructions, featuring a tripartite nominative-accusative-genitive alignment and flexible word order. Beyond the Bible and other religious texts, numerous original and translated works across various genres (such as historiography, mathematics, medicine, etc.) have been preserved in Classical Armenian. Currently, it is maintained exclusively for liturgical purposes.

In contrast, MWA and MEA, and the dialect of Getashen are typologically left-branching, with nominative-accusative alignment and more agglutinative morphology. They exhibit a richer system of periphrastic constructions and maintain flexible word order. MWA and MEA were standardized in the mid-19th century, leading to a rich written heritage. Both standards are currently in use, with MWA primarily by the traditional Armenian diaspora, and MEA used in Armenia, Armenian community of Iran, and Armenian communities in ex-Soviet countries.

The Getashen dialect belongs to the -um branch of the Karabakh dialect family (for more details on Armenian dialects, see Martirosyan (2019) and Davtyan (1966)). This dialect is used in oral form. The data utilized in this study were recorded and transcribed as part of the project "Migration and Complex Identities in the Republic of Armenia (an interdisciplinary study in anthropology and linguistics; Migrant Groups in Armenia (1940-2012): Parameters of Complex Identities)" during fieldworks conducted in Armenia in 2014-2015 (Khurshudyan and Shagoyan, 2016).

3 Related Work

Lemmatization, POS-tagging, and morphological analysis are fundamental tasks in NLP, historically tackled using RNN-based approaches (Manjavacas et al., 2019), or LSTM models often combined with pre-trained word embeddings like GloVe or FastText for contextual word representations. However, state-of-the-art transformer models like BERT (Kondratyuk, 2019), RoBERTa, and XLM-R have significantly outperformed these traditional methods by capturing long-range dependencies and contextual information more effectively through self-attention mechanisms, which allow them to manage complex linguistic patterns.

These transformer-based approaches, though highly effective, generally require large amounts of annotated data, making them less suitable for historical and under-resourced languages due to data scarcity. To address this challenge, recent trends have focused on leveraging attention mechanisms combined with transfer learning from models like DeBERTa (Riemenschneider and Krahn, 2024) or utilizing large language models (LLMs) in assisted annotation frameworks for these languages (Zhao et al., 2024; Bhat and Varma, 2023; Kholodna et al., 2024). Despite these advancements, very under-resourced languages like Coptic, Ancient Egyptian, or Old French still predominantly rely on Seq2Seq architectures, often using LSTM or GRU units with attention mechanisms to handle sequences and generate lemmas or morphological patterns (Manjavacas et al., 2019; Camps et al., 2021; Sahala, 2024).

In the case of Armenian, most lemmatization, POS-tagging, and morphological analysis experiments have focused on Modern Eastern Armenian (Khurshudyan et al., 2022a; Arkhangelskiy et al., 2012) and Classical Armenian (Vidal-Gorène and Kindt, 2020; Kindt and Van Elverdinghe, 2022; Kindt and Vidal-Gorène, 2022; Kharatyan and Kocharov, 2024), using LSTM, joint learning methods with RNNs or rule-based approaches (Khurshudyan et al., 2022b), achieving F1-scores ranging from 0.63 to 0.87 depending on the task and text genre (e.g., Gospel, colophon, HTR output or historiography). These methods have also been applied

	CA	MWA	MEA	G
Tokens	82,557	124,230	52,950	568
Unique tokens	6,837	27,773	14,320	377
Unique lemma	2,472	11,952	7,087	248
Sentences	4,146	6,656	2,500	100
Sentence length (min/max/mean)	2 / 97 / 19.91	1 / 189 / 18.66	2 / 121 / 21.18	27 / 98 / 56.8
Word length (min/max/mean)	1 / 17 / 3.48	1 / 37 / 4.97	1 / 48 / 4.97	1 / 13 / 4.67

Table 1: Overview of the four datasets, including the total number of tokens, unique tokens, unique lemmas, number of sentences, and distributions of sentence and word lengths in defined subsets.

to MEA with similar results (Vidal-Gorène et al., 2020). Experiments in transferring MEA models to Armenian dialects, including MWA, have reported accuracies around 65% in lemmatization and 80% in POS-tagging (Vidal-Gorène et al., 2020).

The application of transformer models or LLMs to Armenian linguistic tasks remains in its early stages, with current usage primarily in classification tasks (Avetisyan et al., 2023).

4 Armenian Datasets

This study draws upon four datasets representing different Armenian dialects. Three of these datasets are sourced from the Universal Dependencies (UD) project (de Marneffe et al., 2021): Classical Armenian (CA)[3], Modern Eastern Armenian (MEA)[4], and Modern Western Armenian (MWA)[5]. The fourth dataset, representing the Getashen (G) dialect, was compiled and transcribed as part of the project "Migration and Complex Identities in the Republic of Armenia" (Khurshudyan and Shagoyan, 2016).

The UD datasets are designed to provide comprehensive morphological and syntactic annotations following UD guidelines, covering a wide range of Armenian language varieties. In contrast, the Getashen dataset consists of raw transcribed text, from which a small number of sentences have been manually annotated specifically for this study.

Modern Eastern Armenian The MWA dataset, also developed by the ArmTDP team, comprises around 52,950 tokens in 2,500 sentences. It spans a wide variety of genres, including blogs, fiction, legal texts, and news. Each sentence is annotated with lemmas, Universal POS-tags (UPOS), and various morphological features, making it the largest manually verified corpus of Eastern Armenian, complete with detailed dependency trees for every sentence.

Modern Western Armenian The MWA dataset, developed by the ArmTDP team, is the most extensive among them, featuring approximately 124,230 tokens across 6,656 sentences, covering a broad range of genres such as blogs, fiction, and nonfiction. The annotation process mirrors that of the MEA dataset, combining automatic glossary-based scripting with manual revision. This dataset is the only manually verified corpus of Western Armenian, offering comprehensive morphological and syntactic annotations.

Classical Armenian The CA dataset is a treebank of the Classical Armenian translation of the four Gospels, by the Classical Armenian-CAVaL treebank project, containing 82,557 tokens in 4,146 sentences. Initially annotated in a non-UD style as part of the PROIEL project, it was later converted to UD format through a rule-based process, followed by manual corrections to ensure accuracy.

Getashen Armenian The fourth dataset, representing the Getashen (G) dialect, consists of a smaller collection of 100 manually annotated sentences. It is used to investigate the transferability of models trained on well-established language variants with long-standing writing traditions and consistent annotation schemas (such as the UD datasets) to a less-documented dialect.

Dataset Statistics Table 1 provides a detailed overview of the composition of these datasets, including statistics on tokens, unique tokens, lemmas, sentences, and the length distributions of both sentences and words. The MEA and MWA datasets, being the largest, show complete alignment in POS-tags, indicating that all POS-tags present in one are also found in the other. They also share the highest overlap in tokens (7.90%) and lemmas (14.25%),

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

	Tokens	Lemmas	POS
MEA-MWA	7.90	14.25	100.00
MEA-CA	2.65	4.77	94.44
MWA-CA	3.31	6.21	94.44
G-MWA	32.36	43.54	36.36
G-MEA	21.22	33.06	36.36
G-CA	11.67	13.70	36.36

Table 2: Percentage overlap (intersection/union) of unique tokens, lemmas, and POS-tags between the four dialect datasets.

suggesting a relatively high degree of linguistic similarity between these two dialects. Table 2 further elaborates on these commonalities, showing that while the MEA-MWA pair exhibits the greatest overlap, the MEA-CA and MWA-CA comparisons have lower overlap in both tokens (2.65% and 3.31%, respectively) and lemmas (4.77% and 6.21%, respectively). This suggests a more distinct linguistic boundary between these datasets.

The Getashen (G) dataset, consisting of transcriptions of spoken language, shows a relatively low overlap with other datasets, ranging from 11.67% to 32.36% for tokens and 13.70% to 43.54% for lemmas. The low overlap in both tokens and lemmas likely reflects the differences inherent in transcriptions of spontaneous speech compared to written text, where greater variability and a broader vocabulary are common. Additionally, the Getashen dataset has an unusually high mean sentence length of 56.8 tokens, contrasting with the shorter averages in the other datasets, which may underscore the complexity and fluidity of spoken language as compared to more structured written forms.

5 Methodology

Our approach aims to understand how different models perform on token-level annotation tasks — lemmatization, POS-tagging, and morphological feature tagging — across multiple Armenian dialects with varying levels of resources and label sets. We explore a unified sequence labeling framework to handle these tasks, leveraging different model architectures, including RNNs, pretrained transformers (mDeBERTa), and large language models (LLMs). By comparing these models in supervised, transfer learning, and zero/few-shot learning settings, we study how well they generalize across dialectal variations and whether combining data from multiple dialects improves performance, particularly for those with limited training data. Codes and raw results are available on Github.[6].

5.1 Task Modeling

The tasks considered in this study — lemmatization, POS-tagging, and morphological feature tagging — are all treated as sequence labeling problems. For each task, a sequence of words (tokens) in a given sentence is mapped to a sequence of labels. *Lemmatization* involves mapping each token to its dictionary form, *POS-tagging* assigns each token its corresponding part-of-speech tag, and *morphological tagging* annotates each token with relevant morphological features (such as case, person, and number).

5.2 Model Architectures

We compare three types of model architectures for the sequence labeling tasks:

Recurrent Neural Network (RNN): An RNN specialized for linguistic tasks (Vidal-Gorène and Kindt, 2020), which builds on has already been used for CA (Vidal-Gorène and Kindt, 2020) and MEA (Vidal-Gorène et al., 2020). Our model relies on the `PIE` architecture (Manjavacas et al., 2019). This method improves annotation of non-standard languages by using an encoder-decoder architecture based on Recurrent Neural Networks (RNNs), enriched with sentence context information through a hierarchical bidirectional RNN and a joint learning approach with a bidirectional language modeling loss. We slightly modify the architecture, adding an attention layer. The RNN models for lemmas, POS-tags and for each morphological feature are trained separately since our prelimenary experiments showed that joint training did not help.

Pretrained Bi-Encoder Transformer: A pretrained mDeBERTa model (He et al., 2021), a multilingual variant of the DeBERTa model, finetuned on the dataset of each dialect. This architecture leverages the power of transformer-based contextual embeddings. Each model consists of the mDeBERTa model, followed by a dropout layer and a linear classifier. Using this setup, the hidden states from the mDeBERTa transformer are mapped to logits that correspond to the labels of each of the

[6]https://github.com/CVidalG/dalih-corpora/

tasks. Models for all tasks share the same backbone transformer and differ only in the classification heads.

Large Language Model (LLM): We employ ChatGPT-4 (OpenAI, 2024) a pretrained large language model in zero-shot and few-shot settings to evaluate its ability to perform the sequence labeling tasks across dialects.

5.3 Learning Paradigms

We explore multiple data setups and learning paradigms to evaluate model performance across different scenarios:

In-Domain Supervised Learning: Each model (RNN and mDeBERTa) is trained in a supervised manner on a specific dialect and evaluated on the same dialect to establish baseline performance.

Cross-Dialect Transfer Learning: To assess the transferability of knowledge, models trained on one dialect are directly evaluated on other dialects without any adaptation. This setup helps us understand how well the models generalize across dialects with different label sets and linguistic characteristics.

Multi-Dialect Supervised Learning: We train the models on the combined datasets of all four varieties to see if pooling data improves performance, especially for dialects with limited training data.

Zero and Few-Shot Learning: We only used ChatGPT-4 in this setup. We aim to evaluate the ability of LLMs to generalize across dialects without explicit training on each. In the few-shot setup, ChatGPT was exposed to a small number of labeled examples using In-Context Learning (ICL) (Brown et al., 2020). We employed three sampling strategies for generation: sequence sampling, random sampling, and less frequent sampling, the latter two strategies performing well in annotation tasks (Bansal and Sharma, 2023). We used sample sizes of 10, 50, 100, and 500 tokens. Experiments were repeated three times, and results were averaged.

5.4 Evaluation Metrics

We use the macro-averaged F1-score instead of the micro-average to give equal weight to all classes, ensuring that the performance on less frequent classes is fairly represented.

	CA	MWA	MEA	G
Lemma				
RNN	0.66	**0.91**	0.79	-
mDeBERTa	0.70	0.44	0.36	-
LLM zero-shot	0.62	0.83	0.74	0.83
LLM in-context	**0.74**	0.83	**0.83**	-
POS				
RNN	**0.98**	**0.98**	**0.98**	-
mDeBERTa	0.91	0.90	0.88	-
LLM zero-shot	0.87	0.86	0.91	0.86
LLM in-context	0.91	0.91	0.85	-
Features				
RNN	**0.88**	0.70	0.66	-
mDeBERTa	**0.88**	**0.78**	0.77	-
LLM zero-shot	0.84	0.71	0.81	-
LLM in-context	0.86	0.75	**0.88**	-

Table 3: F1 macro average results for in-domain supervised learning. The G dialect does not make use of the UD system for features and is not evaluated.

6 Results

6.1 Main Results: Overall Comparison

The results presented in Table 3 show that the RNN consistently performs well across all tasks, particularly for POS-tagging, where it achieves near-perfect scores across the dialects. However, the LLM in-context method often matches or outperforms the RNN for lemmatization and morphological feature tagging, especially in the MEA, indicating its strong adaptability and context understanding. Interestingly, mDeBERTa lags behind in several tasks, particularly for lemmatization in the MEA dialect, suggesting that fine-tuning pretrained models may not always be advantageous compared to both RNNs (specifically designed for the task) and ChatGPT-4. ChatGPT's performance in zero-shot setups also shows its potential for generalization, especially for the G dialect where it performs comparably to supervised methods.

6.2 In-Domain Supervised Learning

We further analyzed the performance of the RNN model which demonstrates strong performance for both lemmatization and POS-tagging on known tokens, achieving high F1-scores across the dialects (e.g., 0.94 for MWA in lemmatization and 0.99 for MWA in POS-tagging). However, its performance significantly drops on unknown tokens, with F1-

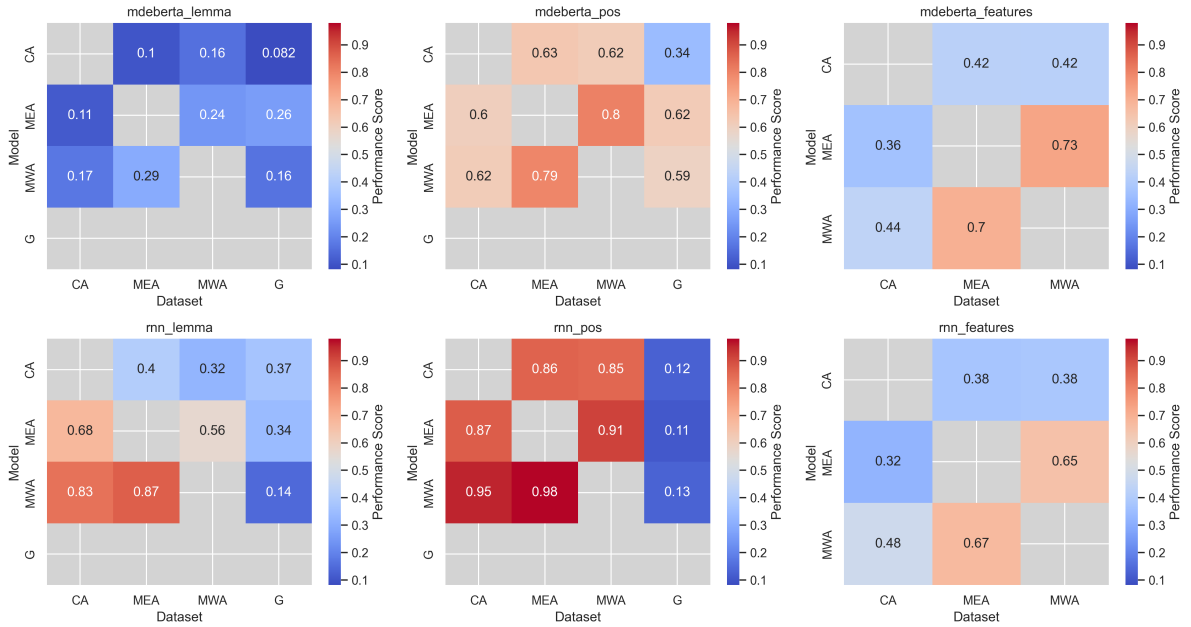

Figure 1: Cross-dialect performance of mDeBERTa and RNN models: Each model was evaluated on datasets outside of its training dialect to assess the generalization capability and immediate applicability without dialect-specific adaptation or mapping. Note that no model is trained on the G dialect as we only have a test set. This G test set contains only lemma and POS annotations.

scores decreasing to 0.43 for MEA and 0.53 for CA in lemmatization, and to 0.50 for CA and 0.53 for MEA in POS-tagging. These results indicate that while the RNN model is effective for known tokens, it struggles with less frequent or unseen classes, particularly in the lemmatization of MEA and CA. In comparison, mDeBERTa does not seem to suffer from this problem, which suggests that combining both models would be beneficial.

While the RNN and mDeBERTa models achieve similar overall performance when averaged across all features, a closer examination reveals that both models excel at handling certain morphological features, such as deixis and tense in MEA, and polarity and person in MWA, with F1-scores near or at 1.0. However, they perform poorly on features related to politeness, degree, and stylistic variations, suggesting that the models are particularly challenged by features that are less frequent or more nuanced in their expression.

6.3 Cross-Variational Transfer Learning

Performance Across Armenian Variation The comparison between mDeBERTa and RNN models across the Armenian dialects (CA, MEA, MWA) highlights the potential and challenges of cross-dialectal modeling for low-resource languages (Figure 1). The RNN consistently outperforms mDeBERTa in lemmatization and POS-tagging, with lemmatization scores ranging from 0.32 to 0.87 and POS-tagging scores from 0.85 to 0.98, compared to mDeBERTa's lower range (0.10 to 0.29 for lemmatization and 0.60 to 0.80 for POS-tagging). However, mDeBERTa performs better on morphological features, achieving scores from 0.36 to 0.73, implying a capacity to handle more generalized tasks despite not being specifically tailored for them.

Dialect Compatibility The results indicate strong compatibility between MEA and MWA for both lemmatization and POS-tagging, reflecting their shared morphological and syntactic structures, with the highest cross-dialect scores at 0.87 and 0.98, respectively. Conversely, the CA dialect shows lower compatibility with modern dialects, particularly in transferring morphological features, where the best CA-to-MWA score is 0.44, pointing to significant linguistic divergence.

Generalization to New Dialects For the new dialect G, although neither model has been specifically trained on it, mDeBERTa and the RNN demonstrate reasonable performance, particularly in POS-tagging and lemmatization (best scores of 0.62 and 0.37, respectively). These findings suggest that cross-lingual transfer and general-purpose models can be valuable for handling linguistic tasks

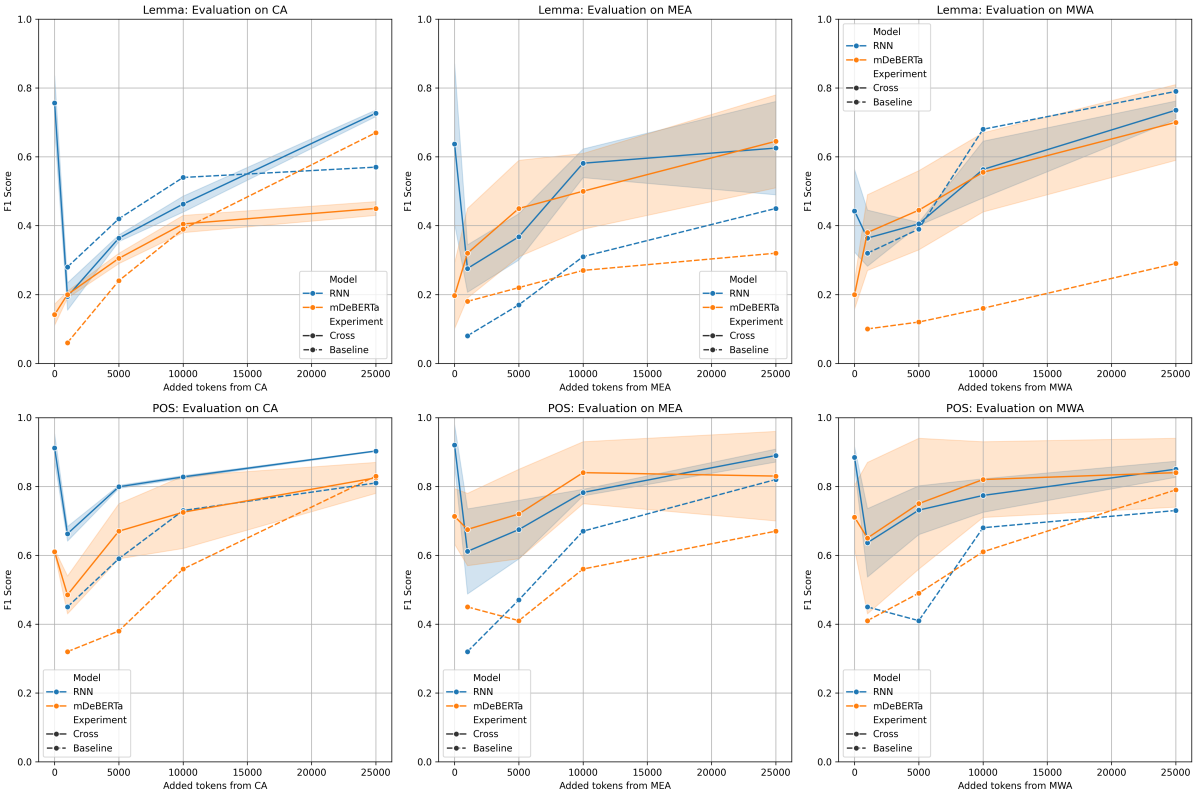

Figure 2: Performance comparison of lemmatization and POS-tagging in multi-dialect supervised learning versus zero-shot transfer learning. Error-bars represent the range of F1-score depending of the used base model (e.g. MWA + CA and MEA + CA for the first subplot)

in under-resourced languages, providing a practical alternative to task-specific models when extensive training data is unavailable.

6.4 Multi-Dialect Supervised Learning

Figure 1 illustrates that adding data from a target dialect to a model trained on a different dialect ("non-specialized model") does not always improve performance. For instance, a non-specialized RNN trained on MEA and evaluated on CA initially achieves a strong F1 score of around 0.68. However, as CA data is incrementally added, the performance of this RNN decreases — dropping to 0.15 with just 1,000 CA tokens. Only after adding 25,000 tokens does the RNN's performance recovers to an F1 score of approximately 0.74, aligning with its initial performance.

In contrast, mDeBERTa, which also starts as a non-specialized model with an F1-score of 0.11 on CA without any CA data, benefits more from adding targeted CA data. By incorporating 10,000 CA tokens, its F1 score rises to 0.43, and with 25,000 tokens, it reaches 0.67, nearly matching the performance of the RNN.

Interestingly, across all evaluation sets (CA, MEA, MWA), non-specialized models (those trained on one dialect and tested on another) often outperform specialized models (those pre-trained and fine-tuned by adding data from the same dialect as the evaluation set). For example, the non-specialized RNN evaluated on MEA without any added MEA data outperforms the specialized RNNs trained directly on MEA, until a significant amount of MEA data is added. This finding highlights the effectiveness of a cross-dialect approach, where training on data from different dialects can lead to better generalization than focusing solely on the target dialect.

6.5 LLM with Few and Zero-Shot Learning

Our goal was to assess how sampling strategy and sample size affect model performance in lemmatization, POS-tagging and full morphological analysis. Evaluations were conducted on a 200-token subset from the test dataset of each language, representing zero-shot performance and varying levels of in-context learning. Results are summarized in Figure 3.

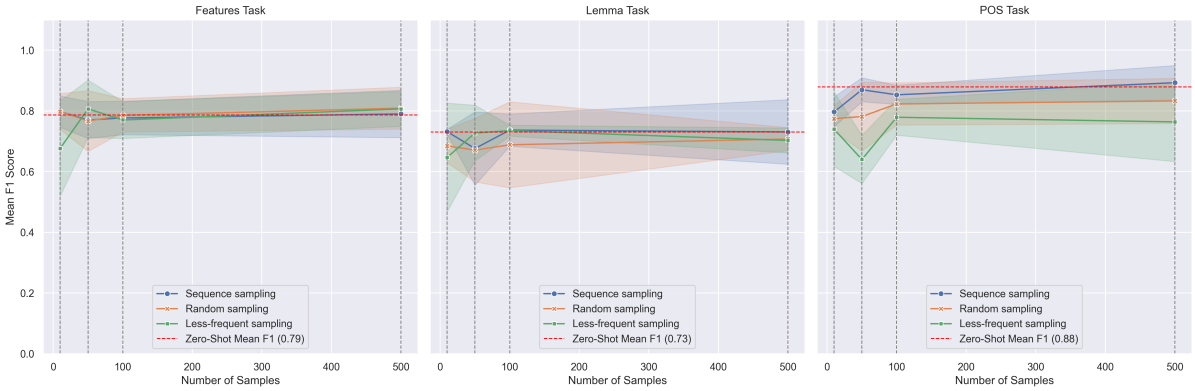

Figure 3: In-context learning using ChatGPT4 with three different sampling strategies: evolution of the mean F1-macro across CA, MEA and MWA.

Overall, the results demonstrate that in-context learning is particularly effective for lemmatization, with sequence sampling consistently outperforming other methods across all tasks and languages. However, for POS-tagging and morphological analysis, the LLM already achieves good results in the zero-shot setting, even for a very under-resourced dialect like G, and further improvements through in-context learning are less pronounced. Sequence sampling works better because it preserves the natural context of tokens, allowing the model to better understand and predict the linguistic patterns within the data. In contrast, random and less frequent sampling disrupts this context, leading to more variability and instability in the results.

Lemmatization The zero-shot F1 score for CA was 0.62, improving to 0.96 with just 10 samples using sequence sampling. For MEA, the zero-shot F1 score was 0.51, which improved significantly with 500 samples, achieving an F1 of 1.0. MWA started with a zero-shot F1 of 0.79, achieving 0.86 with sequence sampling, though additional samples did not consistently lead to improvements. Zero-shot and in-context F1 score on G is similar to MWA.

POS-Tagging POS-tagging using ChatGPT-4 began with a zero-shot F1 score of 0.87 for CA, which remained robust across all sampling strategies. For MEA, the performance improved steadily with sample size, especially with sequence sampling. MWA showed strong initial performance, but additional samples did not yield substantial improvements. Zero-shot and in-context F1 score on G is also similar to MWA.

Morphological Analysis In morphological analysis, sequence sampling led to stable and high F1 scores across all languages, though less frequent sampling exhibited more variability. For CA and MEA, sequence sampling consistently outperformed other methods, while improvements plateaued quickly for MWA.

7 Conclusion

The results from this study demonstrate the varying strengths of different model architectures in handling token-level annotation tasks across multiple Armenian dialects. RNN models consistently performed well, particularly in POS-tagging, where they achieved near-perfect F1 scores of up to 0.98, showcasing their robustness for tasks with known tokens. However, they struggled with less frequent or unseen tokens, where the adaptability of large language models (LLMs), especially in zero-shot and few-shot learning setups, became evident. For instance, ChatGPT-4 in zero-shot learning achieved an F1 score of 0.83 in lemmatization for the Getashen dialect. Pretrained transformers like mDeBERTa, while showing potential in handling morphological features with F1 scores reaching 0.73, often lagged behind RNNs and LLMs, particularly in lemmatization tasks, where their performance dropped to as low as 0.36 in the MEA dialect. Cross-dialect transfer learning revealed that non-specialized models can often generalize better across dialects than specialized ones, suggesting that a cross-dialect approach may be more effective for low-resource languages. In-context learning with LLMs further highlighted their ability to rapidly adapt and improve performance, particularly in lemmatization, where sequence sampling

led to an increase in F1 scores from 0.62 to 0.96 with just 10 samples. In the future, we plan to generalize our approach to include multiple other dialects and to ensure normalization of annotations, facilitating more consistent and comprehensive linguistic analysis.

Perspectives on Interoperability optimization

The annotations applied to the Armenian linguistic data exhibit variability across several dimensions. Firstly, the linguistic level encompasses various types of annotation, including morphological annotation, which involves part-of-speech tagging and the specification of full morphological features, as well as syntactic, semantic, and lexical annotations. Additionally, different categories are utilized to classify various linguistic phenomena, and distinct tagging systems are employed for different features within each annotation type. Moreover, there are notable differences in how morpheme glossing is split. The principles of tokenization are also significant, encompassing considerations such as the presence or absence of spaces and the treatment of internal and external punctuation marks. Finally, the diversity of target language varieties further influences the annotation process.

A potential avenue for further advancement could involve establishing tagging alignment and normalization among the existing datasets. However, automatic mapping without prior analysis and matching is not feasible, as the datasets employ different tagging principles, even though three of them are annotated within the Universal Dependencies framework. A significant systematic issue arises with the MWA and MEA datasets, where formal and functional criteria are mixed, whereas the Classical Armenian dataset employs exclusively formal criteria for tagging.

Another possible approach for dialectal data, for which no written tradition exists, is to process this data through normalization with either the MWA or MEA datasets (Arkhangelskiy and Georgieva, 2018; Waldenfels von R. and Dobrushina, 2014). While this approach may be beneficial for dialectal data, it also necessitates preliminary analysis and specific mapping.

While the aforementioned approaches can lead to significant improvements, establishing a fully harmonized and interoperable annotation system across all projects remains unattainable due to variations in project objectives, linguistic preferences, and the contextual conditions under which these systems were developed. Nevertheless, two parallel pathways can be explored: first, analyzing the existing systems to propose conversion options between them; and second, formulating common principles and annotation solutions for Armenian language data that could be embraced by the user community, while also allowing for conversion into various annotation systems as needed.

8 Acknowledgement

The DALiH project is funded by French National Research Agency ANR-21-CE38-0006.

References

Timofey Arkhangelskiy, Oleg Belyaev, and Arseniy Vydrin. 2012. The creation of large-scale annotated corpora of minority languages using uniparser and the eanc platform. In *Proceedings of COLING 2012: Posters*, pages 83–92.

Timofey Arkhangelskiy and Ekaterina Georgieva. 2018. Sound-aligned corpus of udmurt dialectal texts. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 26–38. Association for Computational Linguistics.

Karen Avetisyan, Arthur Malajyan, Tsolak Ghukasyan, and Arutyun Avetisyan. 2023. A simple and effective method of cross-lingual plagiarism detection. *arXiv preprint arXiv:2304.01352*.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Savita Bhat and Vasudeva Varma. 2023. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 100–107.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jean-Baptiste Camps, Thibault Clérice, Frédéric Duval, Naomi Kanaoka, Ariane Pinche, et al. 2021. Corpus and models for lemmatisation and pos-tagging of old french. *arXiv preprint arXiv:2109.11442*.

Karo Davtyan. 1966. *Leŕnayin Ġarabaġi barbaŕayin k'artezə[= The dialectal map of Nagorno-Karabakh]*. Yerevan.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Anaid Donabedian-Demopoulos. 2018. Middle East and Beyond - Western Armenian at the crossroads : A sociolinguistic and typological sketch. In Christiane Bulut, editor, *A sociolinguistic and typological sketch, in Bulut, Christiane, Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery, , 111/2018, Harrazowitz Verlag*, volume 111 of *Turcologica*, pages 89–148. Harrazowitz Verlag.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DeBERTa}: {Decoding}-{enhanced} {bert} {with} {disentagled} {attention}. In *International Conference on Learning Representations*.

Lilit Kharatyan and Petr Kocharov. 2024. Development of Linguistic Annotation Toolkit for Classical Armenian in SpaCy, Stanza, and UDPipe. In *Proceeding of The First Workshop on Data-driven Approaches to Ancient Languages (DAAL 2024)*, pages 58–66.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. *arXiv preprint arXiv:2404.02261*.

Victoria Khurshudyan, Timofey Arkhangelskiy, Misha Daniel, Vladimir Plungian, Dmitri Levonian, Alex Polyakov, and Sergei Rubakov. 2022a. 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*, pages 28–37, Marseille, France. European Language Resources Association.

Victoria Khurshudyan, Timofey Arkhangelskiy, Misha Daniel, Vladimir Plungian, Dmitri Levonian, Alex Polyakov, and Sergei Rubakov. 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*, pages 28–37.

Victoria Khurshudyan and Anaïd Donabédian. 2021. Cleft constructions and focus strategies in modern armenian. *Faits de Langues*, 52(1):89 – 116.

Victoria Khurshudyan and Gayane Shagoyan. 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*.

Bastien Kindt and Emmanuel Van Elverdinghe. 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*, pages 20–27.

Bastien Kindt and Chahan Vidal-Gorène. 2022. From manuscript to tagged corpora. *Armeniaca-International Journal of Armenian Studies*, 1:73–96.

Dan Kondratyuk. 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*, pages 12–18.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*.

Hrach Martirosyan. 2019. *2.2. The Armenian dialects*, pages 46–105. De Gruyter Mouton, Berlin, Boston.

OpenAI. 2024. Chatgpt-4. `https://openai.com`.

Frederick Riemenschneider and Kevin Krahn. 2024. Heidelberg-boston@ sigtyp 2024 shared task: Enhancing low-resource language analysis with character-aware hierarchical transformers. *arXiv preprint arXiv:2405.20145*.

Aleksi Sahala. 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)*, pages 87–97.

Ollie Sayeed and Bert Vaux. 2017. *66. The evolution of Armenian*, pages 1146–1167. De Gruyter Mouton, Berlin, Boston.

Chahan Vidal-Gorène, Victoria Khurshudyan, and Anaïd Donabédian-Demopoulos. 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).

Chahan Vidal-Gorène and Bastien Kindt. 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.

Daniel M. Waldenfels von R. and N. Dobrushina. 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, pages 720–728.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Appendix

A Detailed Morphological Analysis Results

In MEA, the best-performing features for the RNN model are `deixis[psor]` and `langid` with F1-scores of 1.0, tense at 0.968, `definite` at 0.966, and `case` at 0.952. However, the model performs poorly on features like `polite` (0.333), `number[psor]` (0.333), `person[psor]` (0.249), `degree` (0.243), and `style` (0.124). In MWA, the best features are `polarity` (0.994), `person` (0.990), `tense` (0.988), `definite` (0.987), and `subcat` (0.972). The worst tasks in MWA include `numform` (0.397), `number[psor]` (0.332), `person[psor]` (0.249), `degree` (0.196), and `style` (0.142). These results highlight the model's effectiveness in handling certain morphological features while struggling with others, particularly those involving politeness, degree, and stylistic variations. Additionally, Table 4 presents detailed results for the mDeBERTa model.

B Hyperparameters and Experimental Setup

All hyperparameters, the detailed experimental setup and prompts are released in the accompanying GitHub repository to ensure full reproducibility of the experiments.

Feature	**CA**	**MWA**	**MEA**	**CA > MEA**	**CA > MWA**	**MEA > MWA**	**MEA > CA**	**MWA > CA**	**MWA > MEA**
case	0.96	0.98	0.97	0.62	0.64	0.91	0.71	0.70	0.93
number	0.99	0.97	0.97	0.73	0.72	0.89	0.78	0.77	0.93
person	1.00	1.00	0.99	0.95	0.94	0.96	0.89	0.92	0.98
abbr	-	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
adptype	-	1.00	1.00	0.97	0.96	0.99	0.95	0.93	0.99
animacy	1.00	0.97	0.97	0.71	0.71	0.95	0.81	0.75	0.95
aspect	1.00	1.00	0.99	0.91	0.86	0.95	0.88	0.89	0.95
conjtype	-	1.00	-	1.00	1.00	0.99	0.99	1.00	1.00
connegative	-	1.00	1.00	1.00	1.00	1.00	-	1.00	1.00
definite	1.00	0.98	0.98	0.70	0.70	0.95	0.76	0.77	0.96
degree	-	0.99	0.99	0.96	0.96	0.97	0.99	0.99	0.97
deixis	1.00	1.00	1.00	0.98	0.98	0.99	0.90	0.91	0.99
deixis[psor]	-	1.00	-	-	1.00	-	-	-	-
echo	-	-	1.00	1.00	-	-	-	-	1.00
foreign	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.96	0.99
hyph	-	1.00	-	-	1.00	1.00	-	0.99	1.00
langid	-	-	-	-	-	-	-	-	-
mood	0.99	1.00	0.99	0.95	0.89	0.96	0.91	0.90	0.98
nametype	-	0.98	0.99	0.96	0.96	0.98	0.98	0.98	0.99
number[psor]	-	1.00	1.00	0.99	0.99	1.00	0.98	0.98	1.00
numform	-	1.00	1.00	0.97	0.99	1.00	1.00	0.99	0.98
numtype	1.00	1.00	1.00	0.97	0.99	1.00	1.00	0.99	0.99
person[psor]	-	1.00	1.00	0.99	0.99	1.00	0.98	0.98	0.99
polarity	1.00	0.99	0.99	0.86	0.84	0.96	0.87	0.85	0.96
polite	-	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
poss	1.00	1.00	1.00	0.99	0.98	1.00	0.98	0.99	1.00
prontype	0.99	0.99	0.99	0.95	0.92	0.96	0.85	0.87	0.98
reflex	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99	1.00
style	-	0.99	0.98	0.98	0.99	0.99	1.00	0.99	0.98
subcat	0.99	1.00	0.99	0.92	0.90	0.96	0.91	0.91	0.97
tense	-	1.00	-	-	1.00	1.00	-	-	0.99
typo	1.00	1.00	0.99	0.92	0.94	0.95	0.94	0.95	-
verbform	0.99	0.99	0.98	0.84	0.87	0.97	0.87	0.88	0.95

Table 4: Detailed mDeBERTa morphological analysis results for in-domain supervised learning and cross-dialect transfer learning.