VerbCraft: Morphologically-Aware Armenian Text Generation Using LLMs in Low-Resource Settings

Hayastan Avetisyan and **David Broneske**
German Centre for Higher Education Research and Science Studies (DZHW)
University of Magdeburg
avetisyan@dzhw.eu, broneske@dzhw.eu

Abstract

Understanding and generating morphologically complex verb forms is a critical challenge in Natural Language Processing (NLP), particularly for low-resource languages like Armenian. Armenian's verb morphology encodes multiple layers of grammatical information, such as tense, aspect, mood, voice, person, and number, requiring nuanced computational modeling. We introduce VerbCraft, a novel neural model that integrates explicit morphological classifiers into the mBART-50 architecture. VerbCraft achieves a BLEU score of 0.4899 on test data, compared to the baseline's 0.9975, reflecting its focus on prioritizing morphological precision over fluency. With over 99% accuracy in aspect and voice predictions and robust performance on rare and irregular verb forms, VerbCraft addresses data scarcity through synthetic data generation with human-in-the-loop validation. Beyond Armenian, it offers a scalable framework for morphologically rich, low-resource languages, paving the way for linguistically informed NLP systems and advancing language preservation efforts.

1 Introduction

Armenian, an Indo-European language, presents significant challenges in natural language processing (NLP) due to its intricate verb morphology. Armenian verbs encode multiple layers of grammatical information, including tense, aspect, mood, voice, person, and number, using both synthetic and analytical forms (Dum-Tragut, 2009). This morphological complexity leads to highly nuanced verb forms that are computationally difficult to model.

Morphologically rich languages (MRLs) like Armenian, characterized by their complex inflectional systems and scarcity of annotated data, pose unique challenges for NLP. In such languages, grammatical information is embedded within individual word forms, making accurate modeling essential for tasks such as translation and morphological analysis.

Despite recent advances in neural machine translation (NMT) and pretrained language models, existing approaches often fall short in handling the intricate morphological structures of MRLs. Standard models, such as mBART, struggle to generalize well on low-resource languages, where morphological richness compounds the difficulty of learning effective representations.

To address these challenges, we introduce VerbCraft, a morphologically aware extension of the mBART-50 model. Motivated by the unique morphological complexity of Armenian verbs, VerbCraft incorporates explicit mclassifiers for predicting morphological features into the shared encoder-decoder architecture, bridging the gap between linguistic specificity and translation quality. By explicitly modeling Armenian verb features, such as tense, aspect, and mood, during training, VerbCraft enhances the model's capacity to generate accurate and morphologically consistent translations.

A key feature of this work is the creation of a synthetic dataset using large language models (LLMs), such as ChatGPT, coupled with human-in-the-loop validation by native Armenian speakers. This strategy addresses the scarcity of annotated data for Armenian, enabling the development of robust and linguistically informed NLP models. The dataset includes standard, rare, and irregular verb forms, ensuring comprehensive evaluation of the model's performance.

Through extensive experiments, VerbCraft demonstrates significant improvements over baseline models. Specifically, it achieves a BLEU score of 0.4899 on the test set, compared to the baseline's 0.9975, reflecting its focus on capturing

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 111–119

morphological precision over sentence fluency. In terms of morphological accuracy, VerbCraft consistently outperforms the baseline across key features, achieving 100% accuracy in aspect predictions, 96.26% in voice, 95.33% in tense, and 91.59% in mood. These results underscore the importance of integrating linguistic supervision into NLP systems for morphologically rich languages and highlight the potential for applying this framework to other low-resource languages.

This paper contributes to the field by:

- Introducing **VerbCraft**, a novel neural model integrating morphological classifiers into the mBART-50 architecture, specifically tailored for Armenian verb generation.
- Developing a **synthetic dataset** with ChatGPT and native speaker validation, addressing data scarcity in Armenian NLP.
- Providing a **comparative analysis** demonstrating the advantages of morphologically aware models over traditional sequence-to-sequence models.

This paper is structured as follows: Section 2 reviews related work, highlighting prior efforts in low-resource NLP, NMT, and morphological integration. Section 3 describes the methodology, including model architecture, dataset creation, and evaluation setup. Section 4 presents experimental results and discusses the findings, while Section 5 outlines the limitations of this approach. Finally, Section 6 concludes the paper with insights and directions for future research.

2 Background and Related Work

This section explores prior efforts in integrating morphological features into neural models, particularly for low-resource settings like Armenian, and highlights their applications in neural machine translation and cross-lingual transfer.

2.1 Morphologically Rich Languages in Low-Resource NLP

Languages like Armenian, characterized by complex morphological systems, present significant challenges in NLP due to limited annotated datasets. Morphologically rich languages (MRLs) encode grammatical information, such as tense, aspect, mood, and voice, within individual word forms, resulting in high variability that traditional sequence-based models often fail to capture. Prior works, including KinyaBERT (Nzeyimana and Rubungo, 2022) and MorphoBERT (Mohseni and Tebbifakhr, 2019), demonstrate the value of explicitly integrating morphological features into neural architectures. These studies highlight how morphological information enhances generalization and linguistic understanding in MRLs, especially under low-resource constraints.

Recent studies have also explored the morphological generalization capabilities of LLMs. Dang et al. (2024), for instance, introduced a multilingual adaptation of the Wug Test to assess LLMs' proficiency in applying morphological rules to novel words. Their findings indicate that LLMs can generalize morphological knowledge to unfamiliar terms, with performance influenced by the morphological complexity of each language. Similarly, Ismayilzada et al. (2024) conducted a systematic evaluation of compositional generalization in agglutinative languages like Turkish and Finnish, identifying challenges with novel word roots and increased morphological complexity.

Weller-Di Marco and Fraser (2024) examined LLMs' understanding of morphologically complex German compounds, demonstrating that while LLMs grasp the internal structure of complex words, they often lack formal knowledge of derivational rules, leading to challenges in identifying ill-formed constructions.

Morphological preprocessing techniques, such as those outlined by Straka and Straková (2017), have shown that token-level linguistic features like lemmatization and part-of-speech tagging improve downstream NLP tasks. Additionally, the use of universal dependencies (Nivre et al., 2016) provides a multilingual framework for morphosyntactic analysis, which has inspired methods for integrating rich morphological annotations into neural models.

2.2 Neural Machine Translation and Morphological Features

Neural machine translation (NMT) systems, such as MarianMT and mBART, have been widely adapted for low-resource languages. However, these models often falter when handling extensive morphological variation. Recent approaches, including MorphoBERT and end-to-end lexically constrained NMT (Jon et al., 2021), emphasize the importance of explicitly modeling morphological

features to improve translation accuracy. Arnett and Bergen (2024) discuss how dataset size and tokenization strategies influence performance disparities across typologically diverse languages, underscoring the importance of linguistically informed approaches. Building on these efforts, VerbCraft integrates Armenian-specific morphological classifiers directly into the mBART architecture, enabling precise verb generation and morphological feature prediction.

2.3 LLMs and Data Augmentation

Recent advances in large language models (LLMs), such as GPT-3, offer promising solutions to address data scarcity for low-resource languages. Techniques such as synthetic dataset generation, combined with human-in-the-loop validation, have proven effective for enhancing dataset quality (Santoso et al., 2024). VerbCraft leverages these techniques by employing ChatGPT to generate Armenian verb datasets, which are validated and refined by native speakers. This process ensures linguistic accuracy while addressing the scarcity of annotated resources. Moreover, approaches such as those proposed by Dolatian and Sorensen (Dolatian et al., 2022) provide additional insights into enhancing data generation for underrepresented languages through morphological transducers.

Yin et al. (2024) proposed MorphEval, a benchmark designed to evaluate LLMs' comprehension of Chinese morphemes across characters, words, and sentences. Their evaluation highlights issues such as dysfunctions in morphology and syntax, challenges with long-tailed semantic distributions, and difficulties arising from cultural implications, underscoring the necessity for language-specific enhancements in LLMs. Shin and Kaneko (2024) highlight challenges in modeling character-level information in morphologically complex languages, which are crucial for synthetic dataset creation. Marco and Fraser (2024) further emphasize the role of subword segmentation in improving the recognition and generation of lemmas in morphologically rich languages, aligning with the strategies employed in VerbCraft.

2.4 Cross-Lingual Transfer Learning

Cross-lingual transfer learning provides another avenue for improving NLP tasks in low-resource languages by leveraging data from high-resource counterparts. Methods such as embedding alignment and vocabulary matching (Rybak, 2024) have shown success in tasks like part-of-speech tagging and named entity recognition. Hofmann et al. (2024) investigated linguistic generalization in LLMs, focusing on English adjective nominalization. Their study suggests that LLMs rely more on analogical processes operating on stored exemplars rather than abstract symbolic rules, particularly in cases of variable nominalization patterns.

VerbCraft builds on these ideas by adapting mBART, a multilingual model, for Armenian, explicitly focusing on integrating morphological features. These cross-lingual techniques, combined with recent subword-based methods (Singh et al., 2023), provide a robust foundation for addressing the unique challenges of low-resource morphologically rich languages.

2.5 Research Gap and Contributions

Despite advancements in integrating linguistic features into neural systems, explicit incorporation of explicit classifiers for predicting morphological features for low-resource, morphologically rich languages like Armenian remains underexplored. VerbCraft addresses this gap by embedding Armenian-specific explicit classifiers for predicting morphological features into mBART, demonstrating significant improvements in verb generation accuracy and providing a framework extensible to other MRLs. The alignment with findings from MorphoBERT (Mohseni and Tebbifakhr, 2019) and the emphasis on morphological analysis for downstream tasks (Mohseni and Tebbifakhr, 2019) strengthen its position as a key contribution in this domain. Additionally, insights from Yin et al. (2024) and Beguš et al. (2023) underline the broader necessity of explicit morphological considerations in NLP for low-resource languages.

3 Methodology

This section describes the architecture of VerbCraft, the process of dataset creation, and the evaluation setup, emphasizing the integration of explicit classifiers for predicting morphological features into the mBART-50 model and the strategies used to address data scarcity for Armenian.

3.1 Model Architecture

VerbCraft extends the mBART-50 model by integrating explicit morphological classifiers tailored to Armenian verb morphology. These classifiers predict key grammatical features, including tense,

aspect, mood, voice, person, and number. The architecture is composed of three main components:

1. **Shared Encoder**: The mBART encoder processes the input sequence, generating contextual embeddings that serve as the foundation for both translation and morphological predictions.

2. **Morphological Classifiers**: Separate linear layers are applied to the encoder's embeddings to predict each morphological feature. These classifiers are auxiliary tasks during training, providing additional linguistic supervision and enhancing the encoder's representation.

3. **Decoder**: The decoder generates translations without explicitly incorporating morphological predictions as input tokens, ensuring the sequence-to-sequence nature of mBART is preserved.

The training objective of VerbCraft combines translation and morphological prediction losses to achieve balanced optimization across tasks. Formally, the objective is expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{translation}} + \sum_{f \in \text{features}} \alpha_f \mathcal{L}_f$$

where $\mathcal{L}_{\text{translation}}$ denotes the standard translation loss, and $\mathcal{L}_f$ represents the loss associated with predicting each morphological feature f (e.g., tense, aspect, mood). The weights α_f are empirically tuned to balance the contributions of these auxiliary tasks. This formulation ensures that the model simultaneously learns to generate fluent translations and accurately predict morphological features, enabling it to handle the linguistic complexities of Armenian verbs effectively.

3.2 Dataset

This study employs a novel dataset that encloses the complex morphology of Armenian verbs. The dataset is annotated with fine-grained morphological features, providing a rich resource for NLP tasks focused on Armenian verb generation and analysis.

3.2.1 Dataset Overview and Structure

Our annotated dataset consists of 1,068 sentences and 1,883 annotated verbs, whereby one sentence might encompass more than one annotated verb. Each data point in the dataset is structured as a JSON object containing the following fields:

- `sentence`: The original Armenian sentence.
- `translation`: English translation of the sentence.
- `verb_info`: Detailed information about the verb(s) in the sentence: `tense`, `aspect`, `mood`, `voice`, `person`, `number` and `component breakdown`.

More detailed information on the distribution of morphological features in the dataset can be taken from Table 1.

Tense		**Aspect**	
Aorist	432	Imperfective	1,102
Present	395	Perfective	765
Imperfect	184	Inceptive	15
Future	138	Habitual	1
Conditional	141		
Pluperfect	116		
Present Perfect	112		
Mood		**Voice**	
Indicative	1,266	Active	1,614
Subjunctive	404	Passive	142
Conditional	20	Reflexive	84
Person		**Number**	
3rd Person	1,212	Singular	1,303
1st Person	351	Plural	397
2nd Person	137	None	177
None	177		

Table 1: Distribution of Morphological Features

3.3 Synthetic Data Generation

To address the scarcity of annotated Armenian datasets, we generated synthetic data using ChatGPT [1]. The data generation pipeline includes the following steps:

1. **Prompt Design**: Custom prompts were engineered to produce diverse verb-centric sentences with rich morphological variations. The prompts used for this task can be taken from A.1.

[1] OpenAI, *ChatGPT* (October 2023 version), GPT-4o, 2024, https://openai.com.

2. **Human-in-the-Loop Validation**: Two native Armenian speakers, including a linguist, reviewed and corrected the morphological annotations. This step was crucial to ensure that the dataset reflects linguistic accuracy, especially for irregular verbs or forms with ambiguous meanings.

The final dataset (1,883 annotated verb instances) consists of training, validation, and test splits. An additional inference set (40 instances), enriched with rare and irregular verb forms, evaluates the model's ability to generalize beyond the training distribution.

3.3.1 Preprocessing Steps

The preprocessing pipeline ensures the model receives well-structured input data and correct morphological feature labels. We designed a comprehensive preprocessing function to transform raw input into tokenized sequences and associated morphological annotations.

Tokenization and Feature Extraction:

- **Input Tokenization:** Sentences are tokenized using the *MBart50TokenizerFast* from Hugging Face, which handles multi-lingual text, including Armenian.
- **Morphological Feature Annotation:** Each verb in the input sentence is annotated with its corresponding morphological features. For example, the verb "run" would be encoded as `<VERB:run:<TENSE:past>` to indicate its tense. Additional tags are used for the other features such as aspect, mood, and person.

3.4 Evaluation Setup

We evaluate the system on two main dimensions:

1. **Armenian-to-English Translation**: BLEU scores are computed to measure the fluency and adequacy of the model-generated translations by comparing them with reference translations. This evaluates the model's capability as a translation system.
2. **Morphological Feature Analysis**: Accuracy scores for each morphological feature assess the model's ability to predict explicit linguistic attributes (e.g., tense, aspect, mood) for Armenian verbs. This evaluation highlights the effectiveness of incorporating morphological supervision.
3. **Qualitative Error Analysis**: Qualitative analysis was performed to identify common error patterns, such as tense inconsistencies and incorrect verb conjugations. This analysis provides insights into the model's limitations and guides future improvements.

Additionally, we introduce a specialized inference dataset enriched with rare and irregular verb forms. This dataset is designed to assess the generalization capacity of the model in challenging linguistic scenarios, such as handling verbs with uncommon morphological patterns.

Morphological accuracy is calculated as:

$$\text{Accuracy}_{\text{feature}} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

The evaluation process ensures a comprehensive understanding of the model's strengths and weaknesses, highlighting its ability to handle the complexities of Armenian verb morphology while maintaining translation quality.

3.5 Baseline Model

To contextualize the performance of our enhanced model, we established a baseline using the standard mBART-large-50 model without any morphological enhancements. This baseline serves as a point of comparison, allowing us to quantify the improvements brought about by our architectural modifications and multi-task learning approach. The baseline model was evaluated using the same metrics and datasets as our enhanced model, ensuring a fair and comprehensive comparison.

3.6 Reproducibility

Code, model checkpoints, and datasets are open-sourced to ensure reproducibility. Detailed configuration files for hyperparameters and preprocessed datasets are available as well.

4 Results and Discussion

This section evaluates VerbCraft on the generated dataset, analyzing its performance across various morphological features and translation accuracy.

4.1 Translation Quality: BLEU Score Analysis

VerbCraft's BLEU scores demonstrate a significant improvement in translation quality across epochs:

- Epoch 1: 0.2470

- Epoch 10: 0.4876

However, an anomaly is observed at epoch 5, where the BLEU score temporarily drops to 0.0000 before recovering. This phenomenon likely reflects a shift in internal representations as the model balances translation and auxiliary morphological prediction tasks. Further investigation into these dynamics could optimize learning efficiency.

4.2 Morphological Feature Prediction Accuracy

VerbCraft excels in predicting key morphological features of Armenian verbs during training, as shown in Table 2. The model demonstrates significant improvements across all features, particularly in tense and mood, which are critical for accurate translations.

Feature	Initial Training Accuracy	Final Training Accuracy
Tense	0.0654	0.9813
Aspect	0.0000	1.0000
Mood	0.0000	0.9813
Voice	0.0000	0.9626
Person	0.7103	0.9439
Number	0.0280	0.9626

Table 2: Improvement in Morphological Feature Prediction During Training

These results indicate that the model learned to accurately predict Armenian morphological features during training, crucial for handling the agglutinative nature of Armenian verbs.

A closer look at the learning dynamics of specific morphological features reveals interesting patterns:
Aspect and Voice: These features show rapid improvement, reaching high levels of accuracy early in the training process. Aspect achieves perfect accuracy (1.0000), suggesting that the model fully grasped the distinction between perfective and imperfective forms in Armenian verbs. Similarly, Voice (96.26%) indicates that the model has effectively learned to distinguish between active, passive, and other voice forms.
Tense and Person: The model struggled initially with Tense (0.0654) and Person (0.7103) but showed significant improvement throughout training. The slower improvement may reflect the complexity of the Armenian tense system and agreement patterns requiring more exposure to varied forms in the training data.
Number: The Number feature started with relatively low accuracy (0.0280) but achieved strong performance by the end of training (96.26%). This suggests that singular vs. plural distinctions in Armenian verbs are easier for the model to learn, possibly due to explicit morphological markers in the verb forms.
Mood: The model showed steady improvement in predicting mood (e.g., indicative, subjunctive, imperative), reaching 98.13% accuracy by epoch 10. This suggests that while mood distinctions are challenging, the model can handle them effectively with enough training data and exposure to varied verb forms.

4.3 Comparison of Baseline and Enhanced Model

The comparison between the baseline *mBART-50 model* and the *enhanced VerbCraft* reveals substantial improvements in handling Armenian verb morphology. The enhanced model achieved a BLEU score of 0.4899 on the test set, significantly improving over the baseline model's 0.9975. This improvement reflects the model's ability to generate more syntactically and semantically correct verb forms by effectively capturing complex morphological structures.

Integrating explicit morphological classifiers allowed the enhanced model to outperform the baseline across all key morphological features (see Table 3), particularly in tense and aspect, where accurate predictions are critical. VerbCraft emphasizes morphological precision, possibly at the expense of sentence fluency. This trade-off could lower BLEU scores despite achieving higher accuracy in grammatical features like tense, aspect, and voice. Conversely, the baseline might produce fluent but morphologically inconsistent outputs, inflating BLEU artificially.

The evaluation, conducted on both test and inference sets, showed that the enhanced model demonstrated superior accuracy, confirming that explicitly modeling morphological features leads to significant performance gains in languages with complex verb systems like Armenian.

Metric	Test Data		Inference Data	
	Baseline	Enhanced	Baseline	Enhanced
BLEU Score	0.9975	0.4899	0.9229	0.1060
Tense Acc.	8.41%	95.33%	5.00%	87.50%
Aspect Acc.	39.25%	99.07%	70.00%	100%
Mood Acc.	70.09%	91.59%	87.50%	95.00%
Voice Acc.	81.31%	96.26%	85.00%	92.50%
Person Acc.	14.95%	94.39%	25.00%	97.50%
Number Acc.	78.50%	97.20%	42.50%	100%

Table 3: Performance on Test and Inference Sets

4.4 Error Analysis and Broader Implications

VerbCraft demonstrates notable strengths in handling Armenian verb morphology, while also revealing challenges that highlight broader issues in modeling morphologically rich languages.

4.4.1 Strengths

The enhanced model consistently outperforms the baseline mBART-50 in predicting complex morphological features. Key strengths include:

- **Tense and Person**: VerbCraft excels in predicting morphological features for verbs, particularly in past and imperfect tenses, where the baseline struggled significantly.
- **Aspect and Voice**: With near-perfect accuracy, the model effectively distinguishes between perfective and imperfective aspects, as well as active and passive voice forms.
- **Morphological Awareness**: The ability to process and generate linguistically complex forms demonstrates the model's advanced understanding of Armenian's rich inflectional system.

These strengths underscore the effectiveness of integrating morphological classifiers and linguistic supervision into the model architecture.

4.4.2 Areas for Improvement

Despite its strengths, VerbCraft encounters challenges in balancing grammatical precision with natural language fluency:

- **Tense Consistency**: Errors arise in compound tenses, with occasional mismatches in tense usage within a sentence.
- **Verb Stem Alterations**: Rare but impactful errors involve incorrect modifications of verb stems, altering intended meanings.
- **Auxiliary Verb Omission**: Missing auxiliary verbs in compound tense constructions reduce grammatical completeness.
- **Mood Mismatches**: Generating correct subjunctive and imperative moods remains a challenge, reflecting broader modality modeling issues.

Addressing these issues requires deeper integration of contextual and syntactic information to refine predictions and improve consistency.

4.4.3 Linguistic Insights

The results provide valuable insights into Armenian verb morphology and computational modeling:

- **Aspect and Voice**: Accurate representation of these features is critical for morphologically rich languages and has implications for languages like Turkish and Arabic.
- **Compound Tenses and Mood**: Challenges with auxiliary verb generation and mood predictions highlight the need for nuanced integration of morphology, syntax, and semantics.

4.4.4 Balancing Accuracy and Fluency

The model's high accuracy in predicting linguistic features occasionally comes at the expense of translation fluency. This trade-off reflects the ongoing challenge in NLP for low-resource languages: balancing precise linguistic modeling with coherent and fluent language generation.

4.4.5 Generalization and Broader Relevance

VerbCraft's framework can be adapted to other low-resource, morphologically rich languages such as Finnish, Greek, and Persian. This adaptability offers a roadmap for addressing similar linguistic complexities across diverse languages, advancing NLP for underrepresented linguistic systems.

5 Conclusion and Future Work

VerbCraft successfully integrates explicit morphological classifiers into the mBART-50 framework, addressing key challenges in modeling Armenian verb morphology. The model achieves significant gains in:

1. **Morphological Accuracy**: Achieving 100% accuracy in aspect, 96.26% in voice, 95.33% in tense, and 91.59% in mood predictions, VerbCraft demonstrates its ability to handle the complexities of Armenian verbs.
2. **Morphologically Consistent Translations**: Despite a lower BLEU score (0.4899) compared to the baseline (0.9975), VerbCraft prioritizes grammatical accuracy over fluency, effectively capturing rare and irregular verb forms.

This study establishes a foundation for advancing NLP systems tailored to morphologically rich, low-resource languages. By integrating linguistic supervision into neural architectures, VerbCraft

demonstrates the potential for improving both linguistic precision and translation quality.

Building on these findings, future work will focus on several key areas of improvement and expansion. Firstly, **enhanced contextual modeling** will be explored to address challenges such as tense consistency, auxiliary verb generation, and mood prediction. This will involve incorporating advanced mechanisms to refine the model's contextual understanding.

Secondly, the approach will be extended to include **broader linguistic features**, such as noun morphology and additional dialectal variations. This expansion aims to increase the model's generality and applicability across diverse linguistic contexts.

Thirdly, the methodology will be adapted for **scalability to other languages**, including Greek, Persian, and Turkish. This adaptation will test the framework's potential effectiveness and flexibility in handling diverse linguistic systems.

Additionally, **dataset enrichment** will be prioritized by expanding the current dataset with natural text and multimodal data. This step aims to improve the model's robustness and ability to understand and process richer contextual information.

Finally, future efforts will focus on **integrating syntax and semantics** into the model. By unifying these linguistic layers, the model can achieve holistic linguistic representation, addressing complex phenomena like compound tenses and modal constructions.

Future efforts will address the trade-off between grammatical precision and fluency, optimizing VerbCraft for broader NLP applications while maintaining its focus on linguistic accuracy.

6 Limitations

While VerbCraft represents a significant advancement in morphologically aware NLP for low-resource languages, several limitations warrant attention. VerbCraft faces challenges in balancing accuracy and fluency, with occasional inconsistencies in tense, mood, and auxiliary verb generation. Its reliance on synthetic data and limited dialectal coverage highlight areas for dataset enrichment. Scalability to unrelated languages remains untested, and resource constraints pose practical challenges for widespread adoption. Addressing these issues will refine and generalize the framework further.

A Appendix

A.1 ChatGPT Prompts

Prompt for Dataset Generation: "Generate a diverse set of Armenian sentences with verbs annotated for their morphological features. For each sentence, ensure the verb is annotated with the following features: tense, aspect, mood, voice, person, and number. Include both regular and irregular verbs, as well as a mix of common and rare forms. The output should be formatted in JSON. For each verb, provide: 1) The Armenian sentence. 2) The English translation of the sentence. 3) A detailed breakdown of the verb's morphological features (tense, aspect, mood, voice, person, and number). Generate at least 50 examples featuring verbs across various tenses, aspects, moods, and voices. Ensure the inclusion of sentences containing irregular verbs and complex verb forms, such as the future subjunctive and compound tenses, to capture the full range of Armenian verb morphology." The data was generated between 05.08.2024 and 18.08.2024.


```
{
    "sentence": "Նա գնում էր խանութ:",
    "translation": "He was going to the store."
    "verb_info": {
        "word": "գնում էր",
        "lemma": "գնալ",
        "tense": "imperfect",
        "aspect": "imperfective",
        "mood": "indicative",
        "voice": "active",
        "person": "3",
        "number": "singular",
        "components": [
            {"form": "գնում", "type": "participle"},
            {"form": "էր", "type": "auxiliary", "lemma": "լինել"}
        ]
    }
}
```


Figure 1: Example output (in JSON format).

A.2 Data Split

Number of training samples: 854 Number of validation samples: 107 Number of test samples: 107

References

Catherine Arnett and Benjamin K Bergen. 2024. Why do language models perform worse for morphologically complex languages? *arXiv preprint arXiv:2411.14198*.

Gašper Beguš, Maksymilian Dabkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *arXiv preprint arXiv:2305.00948*.

Anh Dang, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *13th edition of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2024)*, pages 177–188. Association for Computational Linguistics (ACL).

Hossep Dolatian, Daniel Swanson, and Jonathan Washington. 2022. A free/open-source morphological transducer for western armenian. In *Proceedings of the workshop on processing language variation: Digital armenian (DigitAm) within the 13th language resources and evaluation conference*, pages 1–7.

J Dum-Tragut. 2009. Armenian: Modern eastern armenian. amsterdam. *Netherlands, Benjamins*.

Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, and Janet Pierrehumbert. 2024. Derivational morphology reveals analogical generalization in large language models. *arXiv e-prints*, pages arXiv–2411.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Lonneke van der Plas, and Duygu Ataman. 2024. Evaluating morphological compositional generalization in large language models. *arXiv preprint arXiv:2410.12656*.

Josef Jon, João Paulo Aires, Dusan Varis, and Ondřej Bojar. 2021. End-to-end lexically constrained machine translation for morphologically rich languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4019–4033.

Marion Marco and Alexander Fraser. 2024. Subword segmentation in llms: Looking at inflection and consistency. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12050–12060.

Mahdi Mohseni and Amirhossein Tebbifakhr. 2019. Morphobert: A persian ner system with bert and morphological analysis. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, pages 23–30.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363.

Piotr Rybak. 2024. Transferring bert capabilities from high-resource to low-resource languages using vocabulary matching. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16745–16750.

Joan Santoso, Patrick Sutanto, Billy Cahyadi, and Esther Setiawan. 2024. Pushing the limits of low-resource ner using llm artificial data generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9652–9667.

Andrew Shin and Kunitake Kaneko. 2024. Large language models lack understanding of character composition of words. *arXiv preprint arXiv:2405.11357*.

Telem Joyson Singh, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2023. Subwords to word back composition for morphologically rich languages in neural machine translation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 691–700.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 88–99.

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the understanding of morphologically complex words in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1009–1020.

Yaqi Yin, Yue Wang, and Yang Liu. 2024. Chinese morpheme-informed evaluation of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3165–3178.