Building a Clean Bartangi Language Corpus and Training Word Embeddings for Low-Resource Language Modeling

Warda Tariq Victor Popov Vasilii Gromov
Higher School of Economics (HSE University), Moscow, Russia
varda@hse.ru, masterlu@mail.ru, stroller@rambler.ru

Abstract

In this paper, we showcase a comprehensive end-to-end pipeline for creating a superior Bartangi language corpus and using it for training word embeddings. The critically low-resource Pamiri language of Bartangi, which is spoken in Tajikistan, has difficulties such as morphological complexity, orthographic variety, and a lack of data. In order to overcome these obstacles, we gathered a raw corpus of roughly 6,550 phrases, used the Uniparser-Morph-Bartangi morphological analyzer for linguistically accurate lemmatization, and implemented a thorough cleaning procedure to eliminate noise and ensure proper tokenization. The lemmatized corpus that results greatly lowers word sparsity and raises the standard of linguistic analysis. The processed corpus was then used to train two different Word2Vec models, Skip-gram and CBOW, with a vector size of 100, a context window of 5, and a minimum frequency threshold of 1. The resultant word embeddings were displayed using dimensionality reduction techniques like PCA (Pearson, 1901) and t-SNE (van der Maaten and Hinton, 2008), and assessed using intrinsic methods like nearest-neighbor similarity tests. Our tests show that even from tiny datasets, meaningful semantic representations can be obtained by combining informed morphological analysis with clean preprocessing. One of the earliest computational datasets for Bartangi, this resource serves as a vital basis for upcoming NLP tasks, such as language modeling, semantic analysis, and low-resource machine translation. To promote more research in Pamiri and other under-represented languages, we make the corpus, lemmatizer pipeline, and trained embeddings publicly available.

1 Introduction

The Bartangi language is one of the least studied among the Eastern Iranian languages and is spoken by approximately 8,000–10,000 people in the Bartang Valley of the Gorno-Badakhshan Autonomous Region of Tajikistan. It belongs to the Eastern Iranian branch of the Indo-European language family. Bartangi is considered endangered, with no official status and very limited digitized resources. Like many Pamiri languages, it presents significant challenges for computational processing due to its com- plex morphology, rich inflectional system, and orthographic variation.A map of Pamiri language distri- bution (e.g., from Ethnologue or Glottolog) can further illustrate the regional context in which Bartangi is spoken. Low-resource languages are a serious and persistent challenge for the natural language processing (NLP) community (Cotterell and Schütze, 2015). . Despite remarkable achievements in large-scale machine learning and language modeling, computational representation is missing for most of the world's linguistic diversity. Establish- ing fundamental resources such as corpora, lexicons, and word embeddings is essential for supporting basic NLP tasks such as information retrieval, speech processing, and machine translation for these languages (Schnabel et al., 2015). Speakers of such languages risk being left behind in the digital era, restricting their access to modern AI technologies. This study presents a complete pipeline for developing Bartangi language resources from collecting raw corpora to meticulous data cleaning, morphological lemmatization, and word embedding train- ing. We trained Skip-gram and CBOW Word2Vec models (Mikolov et al., 2013) , implemented linguistic normalization using the UniParser-Morph-Bartangi analyzer (Arkhangelskiy, 2019) , and produced a clean, lemmatized Bartangi corpus. The corpus was evaluated quantitatively (via token and lemma statistics) and qualitatively (via nearest- neighbor analysis). We visualized the learned embeddings using PCA (Pearson, 1901) and t-SNE (van der Maaten

Proceedings of Recent Advances in Natural Language Processing,pages 1256–1262
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-145

and Hinton, 2008) to demonstrate their semantic structure. This study supports multilingual AI initiatives (Grave et al., 2018) , helps preserve endangered linguistic resources, and opens the door for further Pamiri NLP research.To encourage reproducibility and continued development, we release our full dataset, code, and trained embeddings publicly at: https://github.com/warda-tariqq/bartangi-language-modeling

2 Related Work

In recent years, the NLP community has shown growing interest in the creation of low-resource language computational resources. Numerous studies have shown that small corpora can provide productive NLP applications for minority languages if cleaned, normalized, and morphologically analyzed properly (Cotterell and Schütze, 2015). For rule-based morphological parsing of complex languages, morphological analyzers like the Uniparser framework have been successful tools. For token-based corpora where every token counts, they support linguistically sensitive lemmatization, which is essential to minimize sparsity.There has been a good precedent of using similar frameworks for Bartangi established by their successful adaptation for languages like Ossetic (Novokshanov, 2021), Avar (Arkhangelskiy, 2020), and Archi (Kibrik and Kodzasov, 2005). Even in situations with little data, Word2Vec has remained a reliable and popular embeddings technique for learning word representations. When trained on carefully selected datasets, Word2Vec models, particularly Skip-gram architectures, can successfully capture semantic links in low-resource languages, as demonstrated by research on Indo-Aryan and Uralic languages (Miyagawa, 2023).Our work expands on existing methods by integrating Word2Vec training and morphological parsing into an integrated pipeline for the Bartangi language. As far as we are aware, this is the first computational attempt to create lemmatized corpora and embeddings for Bartangi, adding a new resource for language processing that is endangered and low-resource.

3 Methodology

Here we describe the complete pipeline developed for the creation of computational resources for the Bartangi language, covering data collection, corpus construction, morphological lemmatization, and word embedding training (Cotterell and Schütze, 2015). Special attention was paid to linguistic accuracy, robustness to noise, and adaptability for downstream NLP tasks, given the low-resource status of Bartangi. The following sections describe each phase of the pipeline in detail.The overall methodology comprises four major phases: data collection, corpus refinement, morphological analysis, and vector representation.The complete pipeline for preparing Bartangi language computational resources is illustrated in Figure 1.

3.1 Data Collection

The first stage in developing Bartangi's computational materials was gathering raw text data. We gathered sentences from a variety of publicly available sources, such as educational books, folklore volumes, and online discussion boards, as there are few digitized materials for Bartangi. Instead of aiming at artificially literary or formal data, the goal was to gather naturalistic linguistic samples representative of conversational language. Following the initial collection, there were roughly 6,550 distinct sentences in the corpus that totaled roughly 25,648 word tokens. Great attention was given to making the data consist of a variety of linguistic events such as postpositions, verb conjugation, and case marking elements all crucial to Bartangi morphology.Because each sentence was saved as an individual.txt file, the corpus structure was simple and modular to retrieve. This structure simplified the subsequent cleaning processes, lemmatization, and analysis in subsequent stages. Parallel processing and reusability were also simplified by this structure. Expansion into the future is also simplified because new sentences can be easily added without affecting the current structure of the corpus. Whereas sentence-level metadata may be included in future corpus iterations, no metadata, such as information regarding speakers or dialect variation, could be included here due to resource limitations on the language.We used the Bartangi corpus publicly released by Novokshanov (Novokshanov, 2020) as the basis for our computational pipeline. This resource consists of manually transcribed sentences from educational materials and folklore sources. We applied additional processing, formatting, and lemmatization scripts (available at our GitHub repository) to prepare it for embedding training and analysis.

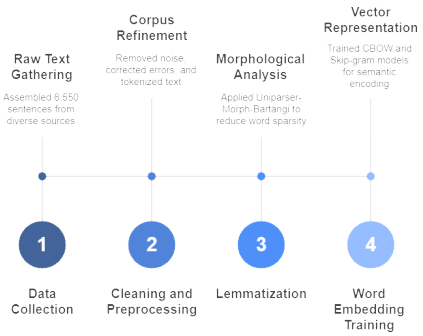

Figure 1: Pipeline for Bartangi Language Computational Resources.

3.2 Cleaning and Preprocessing

For NLP to be successful on low-resource languages, in which noisy or incoherent data can severely degrade model performance, preprocessing of high quality is a condition. In order to handle both orthographic errors and non-linguistic artifacts, we employed a systematic cleaning and preparation procedure in the case of Bartangi.To start, we eliminated unnecessary punctuation that was prevalent in or among words, such as brackets, parentheses, and irregular commas. As a result, tokenization errors (Manning et al., 2008) were minimized and actual linguistic material were kept intact in the text.Essential hyphens and clitics, which contain important morphological information in Bartangi, were kept with utmost caution.Tokenization was performed in the most cautious way so that word boundaries may be maintained even when working with agglutinated forms or morpheme-bound material. Unlike naive whitespace tokenization, our method was capable of identifying cases of misused punctuation on words and splitting them accordingly without breaking semantic consistency.Non-linguistic noise such as markup trash, erroneous characters, and error duplicates were eliminated to avoid subjecting the corpus to proper morphological analysis. One had to avoid including false type lemmas or changing the word type frequency profile.We made sure that the corpus was parseable morphologically, trainable to be embedded, and linguistically correct, clean, and ready through carrying out these pretreatment tasks.

3.3 Lemmatization

Using the Uniparser-Morph-Bartangi system, we performed morphological analysis and lemmatization following cleaning and preprocessing. Lemmatization, which groups inflected, derived, and cliticized word forms into their canonical lemma forms, minimizes the sparsity of surface forms, making it an essential step for low-resource languages.For the purpose of research in this paper, the Uniparser system—which was originally developed for morphologically complex and endangered languages—was specially tuned for the Bartangi language.The analyzer was specifically created to recognize and process the intricate morphological pattern of the Bartangi language accurately. including:

- Prefixes (e.g., verbal prefixes)
- Stems (core lexical roots)
- Suffixes (case endings, verbal inflections)
- Clitics (postpositions, auxiliary markers)

The parser featured grammatical features (i.e., part of speech, case, tense, and number) and recognized the underlying lemma when analyzed. It processed agglutinated forms—words with more than one morpheme attached—properly as well. The outcome was a lemmatized corpus that had a significantly smaller vocabulary, which improved the statistical regularity of the data and the quality of word embedding training.Additionally, we improved the semantic consistency of word vectors learned by lemmatizing before embedding training to prevent semantically identical tokens from being incorrectly modeled as distinct. Overall, the morphological lemmatization step was critical to pre-processing Bartangi language data for downstream machine learning tasks.

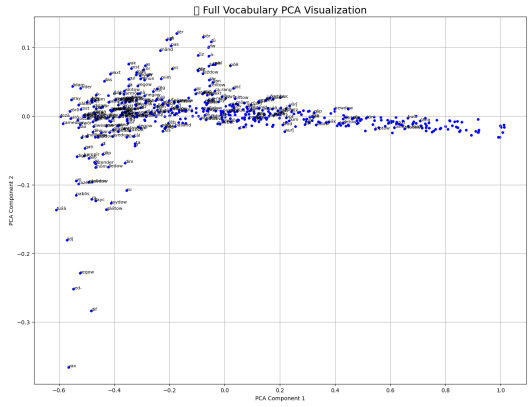

Figure 2: Full vocabulary PCA plot showing semantic groupings of Bartangi lemmas. *Words are shown in Bartangi script; English equivalents are referenced in Table 1.*

3.4 Word Embedding Training

We have learned two Word2Vec models,CBOW and a Skip-gram model, to learn dense vector representations of words in Bartangi. Word embeddings are important for encoding syntactic and semantic regularities between words, particularly when there are limited annotation resources.By making predictions of context words conditional on a target word, the Skip-gram architecture (sg=1) is especially handy for learning effective representations of infrequent or rare words, which is a defining characteristic in small corpora like ours.The CBOW model (sg=0), on the other hand, provides a handy baseline by making predictions of a target word from context words. It also performs marginally better for more frequent tokens. Both models were trained using the Gensim (Řehůřek and Sojka, 2010) library, a commonly adopted toolkit for large-scale NLP modeling (Adelani et al., 2021). Careful selection of training parameters traded off between model capacity and corpus size limits:

- **Vector size**: 100 dimensions
- **Context window size**: 5 tokens
- **Minimum word frequency**(`min_count`): 1

Keeping the low cut-off of the minimum frequency and using a moderate window size ensured low-frequency words in Bartangi also influenced the embedding space. Both Skip-gram and CBOW embeddings were saved to later be used for testing and visual purposes.

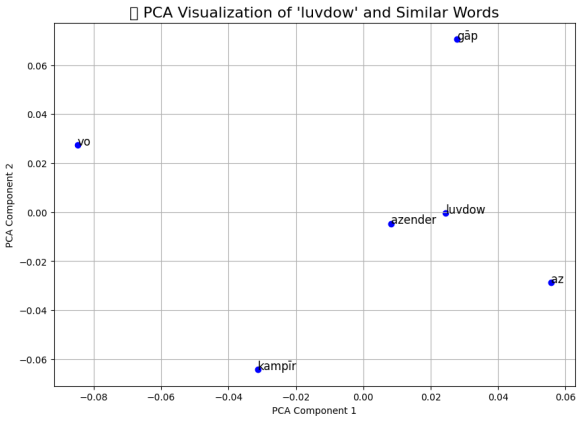

Figure 3: PCA visualization of *luvdow* and its nearest neighbors.

4 Results and Evaluation

Having trained Skip-gram and CBOW models, intrinsic evaluation (Schnabel et al., 2015) was conducted to identify the quality of the word embeddings generated. The vocabulary resulting had about 506 unique words, as would be expected of a cleaned and lemmatized corpus (Alnajjar et al., 2021). Both models were nonetheless capable of generating embeddings that preserved meaningful semantic and syntactic relations among Bartangi words despite the relatively small corpus size.To evaluate the models, we performed nearest-neighbor similarity analysis. For selected target words, we retrieved the top most similar words based on cosine similarity in the embedding space.The embeddings captured verb- and motion-related semantic fields, as demonstrated by the nearest neighbors of the word ”luvdow” (meaning ”to turn” or ”to change”), which included semantically related verbs and grammatical markers such as ”az” (from), ”vo” (and), and ”kampīr” (old woman).In addition, we contrasted how the CBOW and Skip-gram models performed. More dense semantic clusters around rare tokens were formed by the Skip-gram embeddings, which more accurately represented rare and technical words (Miyagawa, 2023). CBOW embeddings were stronger for common words but less sensitive to the subtlety of rare words because they preferred smoothing out common co-occurrences (Hämäläinen et al., 2023).

Model	Bartangi Neighbors	English Gloss
5*Skip-gram	az	from
	azender	approach
	gāp	talk
	vo	and
	kampīr	old woman
5*CBOW	tör	go
	čegow	move
	δedow	come
	az	from
	dif	leave

Table 1: Top-5 Nearest Neighbors for *luvdow* (“to turn”) in Bartangi and their English translations.

Overall, the embeddings proved that vigilant preprocessing and lexically correct lemmatization improve semantic cohesion substantially in word vector spaces when there are paucities of training data.For the purpose of having a clearer insight into the two models’ differences, we compared Skip-gram and CBOW embeddings’ performances on recognizing semantic relations in the Bartangi corpus!(Hämäläinen, 2019).As observed in Table 3,

Statistics	Values
Number of sentences	6,550
Total word tokens	25,648
Unique lemmas	500

Table 2: Corpus Statistics.

Skip-gram embeddings exhibited superior handling of rare and technical terms, which is particularly important owing to the limited size and richness of the Bartangi corpus. CBOW embeddings, however, favored common word patterns but were inferior at handling low-frequency and morphologically complex words (Haddow and Heafield, 2019).

Model	Performance in Your Project
Skip-gram	Better at handling rare and technical words (important because Bartangi is a small, sparse, low-resource corpus). Formed more dense semantic clusters around rare tokens. Preserved context-sensitive information better.
CBOW	Performed better for very common words. Less sensitive to rare words. Smoothed out co-occurrences too much, losing some rare word subtleties.

Table 3: Comparison of Skip-gram and CBOW model performances on Bartangi corpus.

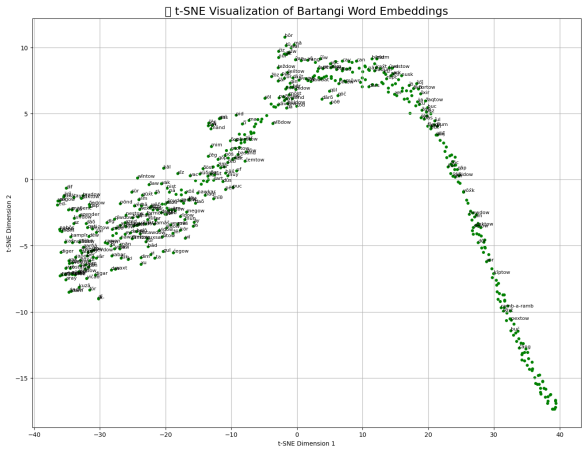

Figure 4: t-SNE visualization of Bartangi word embeddings.
Bartangi forms shown; see Table 1 for English glosses.

5 Visualization

We utilized Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality and plot the learned word vectors in order to evaluate the structure and quality of them qualitatively (Movshovitz-Attias et al., 2020). Both these methods permit qualitative examination of semantic clusters and associations by projecting word vectors in high dimensions onto two dimensions. First, PCA was used to obtain a low-dimensional representation. Separate clusters of semantically similar words were found in the plots.Motion-related, location-related, and object-related words seemed to group around one another, indicating that the embedding space could capture meaningful structures even under a small corpus size. The contextual subtlety sensitivity of the model was seen through the tighter clustering of related words shown by Skip-gram embeddings in particular (Thompson and Saranpää, 2021).

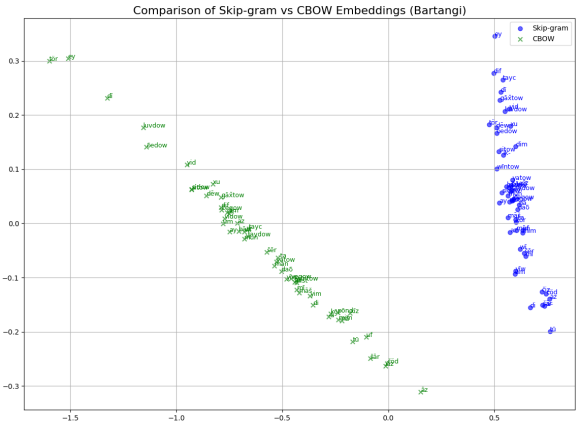

Figure 5: Skip-gram vs. CBOW embeddings (Bartangi). *Trained independently; not aligned in same vector space.*

6 Conclusion

In this paper, we introduce a complete pipeline for constructing Bartangi language resources, including training word embeddings, morphological lemmatization, corpus cleaning, raw data collection, and evaluation. We can observe from our experiment that rigorous model training and heavy linguistic preprocessing can successfully produce meaningful computational representations, even for very low-resource languages with rich morphology and little textual material.We created a linguistically accurate lemmatized corpus using the

Uniparser-Morph-Bartangi morphological parser, which largely alleviated sparsity of vocabulary and enhanced data quality. Both internal evaluation and qualitative visualization demonstrate that Skip-gram and CBOW Word2Vec embeddings created by our method reflect meaningful semantic and syntactic relationships between Bartangi words.This dataset is among the publicly released NLP resources for Bartangi, opening up avenues for further research on Pamiri endangered languages as well as more general multilingual NLP projects.We make the processed Bartangi corpus, morphological lemmatization pipeline, and trained word embeddings (Skip-gram and CBOW) publicly available to facilitate future research on low-resource languages. Future project directions include training more sophisticated embedding models like FastText, which is better equipped to handle rare wordforms, or contextual embeddings like BERT-based models fine-tuned on small corpora; incorporating grammatical metadata into the annotations; and increasing the corpus size through the inclusion of more raw textual materials. Future project directions include training more sophisticated embedding models like FastText, which is better equipped to handle rare wordforms, or contextual embeddings like BERT-based models fine-tuned on small corpora; incorporating grammatical metadata into the annotations; and increasing the corpus size through the inclusion of more raw textual materials.All other things being equal, this study shows that significant steps toward computational modeling of under-documented languages are possible even from small, well-designed datasets.

References

David Ifeoluwa Adelani et al. 2021. A few thousand translations go a long way: Leveraging pretrained models for african news translation. In *Findings of ACL*. Association for Computational Linguistics.

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2021. When word embeddings become endangered: Cross-lingual embeddings for erzya, moksha, komi-zyrian & skolt sami. *arXiv preprint*.

Timofey Arkhangelskiy. 2019. Uniparser: A rule-based morphological parser.

Timofey Arkhangelskiy. 2020. Developing a rule-based morphological analyzer for the avar language. In *Proceedings of the Workshop on NLP for Caucasian Languages*.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of NAACL*. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*. European Language Resources Association.

Barry Haddow and Kenneth Heafield. 2019. Bee language technology for endangered celtic languages: A case study on cornish. In *Proceedings of the 13th International Workshop on Computational Linguistics for Uralic and Altaic Languages*, Tartu, Estonia.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for uralic languages. *Journal of Open Source Software*, 4(37):1345.

Mika Hämäläinen, Jack Rueter, Khalid Alnajjar, and Niko Partanen. 2023. Working towards digital documentation of uralic languages with open-source tools and modern NLP methods. In *Proceedings of the Big Picture Workshop*, Singapore. Association for Computational Linguistics.

Andrej Kibrik and Sandro Kodzasov. 2005. *The Archi Language: Morphological and Syntactic Features*. Languages of the World.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Foundations of Statistical Natural Language Processing*. MIT Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the Workshop at ICLR*. International Conference on Learning Representations.

So Miyagawa. 2023. Machine translation for highly low-resource language: A case study of ainu, a critically endangered indigenous language in northern japan. In *Proceedings of the Joint 3rd International Conference on NLP for Digital Humanities & 8th IWCLUL*, Tokyo.

Roy Movshovitz-Attias, Aaron Smart, and Theodore J. Liu. 2020. Neural sequence tagging for very low-resource languages with application to biblical hebrew. In *Findings of EMNLP*, pages 2430–2442. Association for Computational Linguistics.

Arseniy Novokshanov. 2020. Bartangi corpus (tsakorpus project).

Arseniy Novokshanov. 2021. A morphological analyzer for ossetic: Design, implementation, and evaluation. In *Proceedings of the International Workshop on Computational Morphology*.

Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*. European Language Resources Association.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Lucy Thompson and Kaisa Saranpää. 2021. Neural machine translation for skolt sami: A critical low-resource scenario. In *Proceedings of the 16th Conference on Language Resources and Evaluation (LREC)*, Marseille. European Language Resources Association.