Subword Tokenization Strategies for Kurdish Word Embeddings

Anonymous Authors[1]

Abstract

We investigate tokenization strategies for Kurdish word embeddings by comparing word-level, morpheme-based, and BPE approaches on morphological similarity preservation tasks. We develop a BiLSTM-CRF morphological segmenter using bootstrapped training from minimal manual annotation and evaluate Word2Vec embeddings across comprehensive metrics including similarity preservation, clustering quality, and semantic organization. Our analysis reveals critical evaluation biases in tokenization comparison. While BPE initially appears superior in morphological similarity, it evaluates only 28.6% of test cases compared to 68.7% for morpheme model, creating artificial performance inflation. When assessed comprehensively, morpheme-based tokenization demonstrates superior embedding space organization, better semantic neighborhood structure, and more balanced coverage across morphological complexity levels. These findings highlight the importance of coverage-aware evaluation in low-resource language processing and offers different tokenization methods for low-resourced language processing.

1. Introduction

Effective word representations are critical for natural language processing, particularly for low-resource, morphologically rich languages where data scarcity compounds the complexity of linguistic structures (Erdmann & Habash, 2018; Ruder et al., 2019). Tokenization, the process of segmenting text into units, is the foundation for these representations, with significant impact on downstream applications. Conventional word-level tokenization approaches fail to capture the compositional nature of meaning encoded in morphological structures in languages with rich morphology (Cotterell & Schütze, 2016). While linguistic theory suggests morpheme-based tokenization should outperform statistical approaches for morphologically rich languages (Luong et al., 2013; Park et al., 2020), some empirical evidence has begun to challenge this assumption. Recent research across various languages has revealed a surprising gap between theoretical expectations and practical results in subword tokenization (Bostrom & Durrett, 2020). Statistical methods like Byte-Pair Encoding (BPE) (Sennrich et al., 2016), which merge frequent character sequences without linguistic guidance, sometimes outperform linguistically-informed approaches (Mielke et al., 2019). On the other hand, comparative research on morphologically rich languages like Turkish, Finnish, and Hungarian (Cotterell & Schütze, 2016; Creutz & Lagus, 2005) has demonstrated the benefits of morphologically-informed representations.

In this work, we present an in-depth analysis of the importance of tokenization for Kurdish natural language processing due to its complex morphology (Esmaili & Salavati, 2013). Despite its historical and cultural significance, Kurdish remains under-resourced in computational linguistics (Hassani, 2018), with limited exploration of optimal tokenization strategies. The unexpected findings that subword tokenization is sufficient for most tasks necessitates systematic evaluation of different segmentation strategies. Here we ask which tokenization approaches are most optimal for Kurdish to narrow the gap in the literature (Ahmadi & Wurm, 2019). We review the linguistic properties of Kurdish and present a comparison of word-level, morpheme-based, and subword tokenization strategies for Kurdish, examining their impact on word embedding quality and downstream tasks. Our work addresses the crucial need for empirically-grounded tokenization approaches in low-resource settings (Gerz et al., 2018) and contributes to the broader understanding of representation learning for morphologically complex languages.

2. Challenges for Kurdish Language Processing

Kurdish morphology exhibits extensive derivational and inflectional processes (Thackston, 2006). Nouns encode definiteness and number, while verbs express tense, aspect, mood, person, and number through complex affixation patterns. The agglutinative structure of Kurdish partic-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ularly complicates verbal constructions where multiple morphemes concatenate to form complex predicates (Ahmadi et al., 2020). For example, the Kurdish verb *neyandegerandewe* (they wouldn't return it) contains six morphemes: *ne-* (negation) + *yan* (3rd person plural object) + *de-* (past imperfective) + *gerênd* (causative 'return') + *ewe* (directional 'back'). The morphological density of Kurdish words is comparable to other morphologically rich languages and creates compositional structures where single words encode multiple features. The noun کتێبەکانم (*kitêbekanim*, "my books") demonstrates typical complexity: کتێب (*kitêb*, "book") + ەکان (*ekan*, plural) + م (*im*, possessive). This morphological complexity yields high type-to-token ratios, exacerbating data sparsity in computational models.

In addition to its morphological properties, Kurdish lacks the large-scale annotated corpora and computational lexicons available for high-resource languages (Hassani, 2018), though large, unlabeled datasets are increasingly available. Dialectal variation and sociolinguistic fragmentation adds complexity, with Sorani and Kurmanji differing in morphology, syntax, and vocabulary (Sheykh Esmaili & Salavati, 2013). Orthographic inconsistencies arise from optional vowel representation in the Arabic-derived script, where short vowels may be omitted (Ahmadi, 2020). Morphological analysis tools have developed incrementally, with notable efforts including Sorani analyzers (Ahmadi & Hassani, 2020; Ahmadi, 2021), finite-state transducers for Kurdish (Ahmadi & Hassani, 2020), and the comprehensive AsoSoft framework (Veisi et al., 2019; Mahmudi & Veisi, 2021), which provides large text collections, transliteration systems, and web-accessible morphological analysis. These tools achieve reasonable coverage for standard texts but struggle with neologisms, borrowed terminology, and social media content with spelling variations.

We argue here that linguistically informed tokenization that combines the strengths of statistical methods like subword segmentation (Sennrich et al., 2016) and morphological boundary detection presents a potential solution for the orthographic and morphophonemic alternations of Kurdish. To address these issues, we present a neural BiLSTM-CRF model that addresses these limitations through bootstrapped training from minimal annotation (1,540 words), demonstrating effective morphological analysis with limited resources.

3. Methodology

The training process for our morphological segmentation model followed a bootstrapping approach. Our study utilized the AsoSoft Text Corpus (Veisi et al., 2019) as the primary data source, which is one of the largest available collections of Kurdish (Sorani) text. We manually segmented approximately 1,500 Kurdish words based on linguistic morphological analysis. These words were randomly selected from the corpus to ensure coverage of different word types and morphological patterns including different parts of speech, light verb constructions, preverbal constructions and compounds. These words served as the initial training set for the BiLSTM-CRF model. After training this initial model, we applied it to segment additional words from our corpus, manually verified a subset of these new segmentations, and added them to our training data. Through this iterative process, we expanded our training set to over 4,000 words with gold-standard morphological segmentations.

3.1. Text normalization

Text normalization for Kurdish presented numerous challenges due to its non-standardized orthography, dialectal diversity, and the nature of the available corpus. Given the lack of standard Kurdish NLP preprocessing libraries, every aspect of the workflow had to be developed from scratch or extensively adapted. Our preprocessing pipeline was developed across several stages, requiring considerable manual tuning and verification. We applied AsoSoft's text normalization method (Mahmudi & Veisi, 2021; Veisi et al., 2019) from their Python library in the initial preprocessing stages to address standard Kurdish text inconsistencies before implementing our custom normalization procedures. This multi-phase normalization required several iterations of testing, reviewing, and modifying rules across millions of tokens. The final preprocessed corpus formed the foundation of our tokenization and embedding experiments, and its quality was critical to the validity of all subsequent results.

The Asosoft corpus is constructed from various sources, with news articles constituting the majority of the content. After extensive preprocessing, our final cleaned corpus contained 24.5 million tokens spanning approximately 2.3 million sentences. We first defined a strict set of allowable characters that included letters from the extended Arabic-based script used in Sorani Kurdish, numerals, and a limited range of punctuation marks. All non-Kurdish or extraneous characters were filtered using regular expressions. This filtering process was not straightforward, as informal writing styles, character borrowing from Persian and Arabic, and inconsistent Unicode encodings introduced significant noise in the text (see Appendix).

Corpus cleanup also included extensive deduplication, including exact sentence duplicates and fuzzy duplicates using token-overlap measures. We removed near-identical headlines, repeated paragraphs, and templated sentences across documents. Sentences that fell below a minimum token threshold or lacked valid word structure were discarded (see appendix). In the final phase of preprocessing, we implemented Kurdish-specific sentence segmentation rules to extract clean sentence boundaries using heuristics around punctuation and spacing. Given the inconsistent use of sentence delimiters and overlap with non-Kurdish scripts, this step required custom filtering to remove embedded Persian and Arabic segments. The resulting corpus of 2.3 million well-formed sentences was further deduplicated and formatted into sentence-per-line and word-per-line variants to support downstream tokenization schemes.

3.2. Tokenization approaches

Tokenization strategies for morphologically rich languages fundamentally shape how models represent linguistic structure, spanning from statistical to linguistically-informed methods. This comparison addresses a core theoretical question: whether linguistically-informed segmentation outperforms statistical frequency-based approaches for capturing meaningful morphological relationships in Kurdish. We quantify the convergence between these approaches through segmentation agreement analysis, measuring boundary alignment using similarity coefficients.

Statistical approaches like Byte-Pair Encoding (BPE; Sennrich et al., 2016) operate through iterative merging, starting with characters and incrementally merging frequent adjacent pairs $p = (x, y)$ using $\arg\max_{p \in V} \text{count}(p)$ until reaching target vocabulary size. Alternative methods include WordPiece (Schuster & Nakajima, 2012), which incorporates likelihood criteria, and SentencePiece (Kudo & Richardson, 2018), which treats whitespace as regular characters. The unigram language model (Kudo, 2018) employs a different paradigm, starting with large vocabulary and iteratively removing subwords to maximize corpus likelihood. Unsupervised morphological approaches include Morfessor (Smit et al., 2014), which applies minimum description length principles to automatically discover morpheme boundaries. This method seeks segmentations that minimize combined encoding costs of both lexicon (morpheme inventory) and corpus (word occurrences), balancing between poor corpus compression (too few morphemes) and excessive lexicon size (too many morphemes).

By contrast, linguistically-motivated segmentation identifies meaningful units s aligned with linguistic structures such as words (w) rather than frequency patterns, formulated as:

$$\arg \max_{s_1, s_2, \ldots, s_n} P(s_1, s_2, \ldots, s_n | w) \tag{1}$$

where the probability function incorporates morphological knowledge about valid combinations and morphotactic constraints.

3.2.1. Morphological segmentation with BiLSTM-CRF

For morpheme-level tokenization, we developed a BiLSTM-CRF neural architecture (Lafferty et al., 2001; Huang et al., 2015) that predicts morpheme boundaries within Kurdish words. This approach decomposes complex forms into constituent morphemes, enabling capture of morphological regularities and improved generalization across paradigms. Given Kurdish's limited annotated data, we investigate whether effective segmentation is achievable through bootstrapping from minimal training data.

3.2.2. Model Architecture

The BiLSTM-CRF model processes words at the character level using an embedding-based approach. Our implementation begins with a character embedding layer that maps each character to a dense vector representation:

$$\mathbf{x}_t = \text{Embedding}(c_t) \in \mathbb{R}^d \tag{2}$$

where d is the embedding dimension for each character in the Kurdish alphabet. The architecture consists of three main components: a multi-layer bidirectional LSTM that processes character sequences in both directions to capture contextual information, a linear projection layer that maps LSTM outputs to boundary prediction scores, and a Conditional Random Field (CRF) layer that enforces valid boundary label sequences.The model computation proceeds as follows:

$$\mathbf{h}^f, \mathbf{h}^b = \text{BiLSTM}(\mathbf{x}_{1:n}) \tag{3}$$

$$\mathbf{h}_t = [\mathbf{h}_t^f; \mathbf{h}_t^b] \tag{4}$$

$$\mathbf{e}_t = \mathbf{W}\mathbf{h}_t + \mathbf{b} \tag{5}$$

The CRF layer models the conditional probability of label sequence $\mathbf{y}$ given input $\mathbf{x}$:

$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp(s(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{x})} \exp(s(\mathbf{x}, \mathbf{y}'))} \tag{6}$$

where the score function combines emission and transition scores:

$$s(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{n} [\mathbf{e}_{t, y_t} + \mathbf{T}_{y_{t-1}, y_t}] \tag{7}$$

Here, $\mathbf{h}^f$ and $\mathbf{h}^b$ represent the forward and backward LSTM hidden states, $\mathbf{e}_t$ are emission scores, and $\mathbf{T}$ is the transition score matrix.

3.2.3. Training Configuration and Bootstrapping

We employed a bootstrapping approach to train the model with minimal annotation resources. Beginning with 1,540 manually segmented Kurdish words, we iteratively expanded our training set through model-assisted annotation and manual verification, ultimately reaching over 4,000 annotated words. The model used a hidden size of 256 dimensions, 3 BiLSTM layers, and a dropout rate of 0.3 for regularization. We used the Adam optimizer with learning rate 0.001 and weight decay 1e-5, employing early stopping with 10 epochs patience to prevent overfitting.

We experimented with two labeling schemes: the "end-only scheme" where only the last character of each morpheme is marked as a boundary (0-0-1), and the "both-ends scheme" where both the first and last characters of each morpheme are marked as boundaries (1-0-1). After evaluation on our validation set, we selected the end-only scheme for our final model based on its superior performance. During inference, the model segments words using the CRF's Viterbi decoding algorithm (Lafferty et al., 2001) to find the optimal label sequence:

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}(\mathbf{x})}{\arg\max} \sum_{t=1}^{n} [\mathbf{e}_{t,y_t} + \mathbf{T}_{y_{t-1},y_t}] \tag{8}$$

3.2.4. Model Performance and Error Analysis

The final model achieved an F1-score of .815 for boundary detection, with precision of .835 and recall of .796. However, performance evaluation revealed significant variation across word categories. Nouns achieved over 90% segmentation accuracy and adjectives around 93%, while verbs reached only 42% accuracy. This discrepancy stems from Kurdish verb morphology's greater complexity, including extensive affixation patterns, irregular forms, single-character stems with phonological changes, ambiguous direction-marking affixes, and fewer training examples of complex verbal constructions.

After applying morphological segmentation to our full corpus, each word decomposed into an average of 1.99 morphemes, consistent with expectations for morphologically complex languages. This segmentation strategy significantly impacts embedding quality for morphologically related forms, particularly for nouns and adjectives where segmentation accuracy is high. We examine how this segmentation quality directly affects morphological similarity preservation in our embedding evaluation results (Section 4.4).

3.2.5. Byte-Pair Encoding (BPE)

We implemented Byte-Pair Encoding (Sennrich et al., 2016) as a tokenization approach that operates without explicit linguistic knowledge. BPE offers a data-driven alternative to morphologically-informed segmentation.

Our BPE implementation used the HuggingFace Tokenizers library with a target vocabulary size of 2,280 tokens based on empirical morpheme count analysis. This vocabulary size was chosen to create a compact representation that balances coverage and efficiency—small enough to ensure frequent occurrence of each subword unit for robust statistical learning, yet large enough to capture common Kurdish character sequences and morphological patterns. The training process included a minimum frequency threshold of 2 to filter out rare character combinations.

The resulting BPE tokenizer segments Kurdish words into subword units averaging 3.75 tokens per word, more granular than the morpheme-based approach (1.99 tokens per word) but more compact than character-level tokenization. For example, a complex Kurdish word like دەستپێکردنەوە ("to start") might be segmented as دەست_پێ_کرد_نەوە, where the algorithm has learned to identify frequently co-occurring character sequences regardless of their linguistic significance.

3.2.6. Word-level tokenization

Word-level tokenization serves as our baseline approach, treating each word as a unit without internal decomposition. This traditional method employs whitespace and punctuation-based segmentation, creating the largest vocabulary among our three approaches. For Kurdish text processing, we applied standard tokenization rules while addressing language-specific considerations such as compound word boundaries and clitic attachment patterns.

While word-level tokenization offers the advantage of preserving complete lexical meanings, it presents significant challenges for morphologically rich languages like Kurdish. The approach suffers from high data sparsity, as each inflected or derived form is treated as a distinct vocabulary item, and cannot generalize to unseen word forms. This limitation is particularly pronounced in Kurdish, where productive morphological processes can generate numerous variants of a base form, leading to substantial out-of-vocabulary issues and requiring extensive vocabulary coverage for adequate representation.

3.3. Kurdish word, subword, and morpheme embeddings

We employed the skip-gram word2vec architecture (Mikolov et al., 2013) to train comparable embeddings across all three tokenization approaches. Our framework addresses the challenge of fair comparison between tokenization strategies that produce vastly different token counts per word. Our evaluation includes analysis of embedding space organization through separation ratios (measuring intra-lemma vs. inter-lemma clustering) and similarity dropoff rates (quantifying how similarity decreases across nearest neighbor rankings).

3.3.1. Window Size Adjustment Methodology

A critical part in our approach is the dynamic adjustment of context window sizes based on tokenization granularity. Since different tokenization strategies produce varying numbers of tokens per word, maintaining identical window sizes would create unfair comparisons where subword approaches see artificially truncated contexts. We calculate the average tokens per word for each tokenization approach and adjust window sizes proportionally:

$$w_{adjusted} = \lceil w_{base} \times \text{avg\_tokens}_{approach} \rceil \tag{9}$$

where w_{base} represents the baseline window size for word-level tokenization.

3.3.2. Training Configuration

All models used identical training parameters: vector dimensions of 100 and 150, minimum count threshold of 5, 10 training epochs, and skip-gram with negative sampling (5 negative samples). Models were trained using Gensim's word2vec (Řehůřek & Sojka, 2010) implementation with consistent preprocessing pipelines to eliminate confounding factors. For out-of-vocabulary evaluation, morpheme and BPE models compute compositional vectors by averaging constituent subword embeddings, while the word-level model cannot handle unseen words. Training consistency was ensured through fixed random seeds and multiple independent runs, with final results representing averaged performance across three training iterations.

3.3.3. Handling Out-of-Vocabulary Words

A critical consideration for fair evaluation is handling words not present in model vocabularies during testing. The three tokenization approaches exhibit fundamentally different capabilities for addressing out-of-vocabulary (OOV) words, which directly impacts evaluation coverage and comparability. Coverage refers to the percentage of evaluation word pairs for which a model can generate embeddings for both components. We utilize coverage as the percentage of evaluation word pairs for which a model can generate embeddings for both the lemma and wordform, either through direct vocabulary lookup or compositional vector construction. Compositional vectors are created by averaging embeddings of constituent subword units when complete words are absent from the vocabulary. Compositional vectors are constructed by averaging the embeddings of constituent subword units when the complete word is not present in the model vocabulary.

For morpheme and BPE models, we implement compositional vector generation by averaging constituent subword embeddings when the complete word form is absent from the vocabulary. When evaluating a word not present in the morpheme model, we decompose it using our BiLSTM-CRF segmenter and compute the mean vector of available morpheme embeddings. Similarly, BPE models leverage their learned subword units to construct representations for unseen words through vector averaging of constituent BPE tokens. The word-level model, by design, cannot generate representations for OOV words, as it treats each word as an atomic unit. This limitation creates an inherent evaluation disadvantage, as word-level models have zero coverage for words absent from their training vocabulary.

4. Evaluation and Results

Our evaluation methodology implements a framework designed to establish the relationship between segmentation quality and embedding performance while identifying distinct organizational patterns that different tokenization strategies create in embedding space. We present evaluation methodology alongside corresponding results to provide immediate insights into each analytical approach.

4.1. Morphological Segmentation Quality Assessment

We employ the UniMorph Kurdish dataset (Pimentel et al., 2021) as our gold standard for morphological segmentation evaluation. UniMorph provides morphologically annotated word forms with lemmas and feature annotations for about 1,000 Kurdish (Sorani) words, making it suitable for validating our BiLSTM-CRF segmentation accuracy. We evaluate the BiLSTM-CRF segmentation system using F1-score calculation for binary boundary detection, treating each character position as either boundary (1) or non-boundary (0). Performance assessment includes systematic breakdown by part-of-speech categories, revealing substantial variation in segmentation difficulty across linguistic categories.

Figure 1 shows the substantial performance variation across POS categories detailed in our methodology.

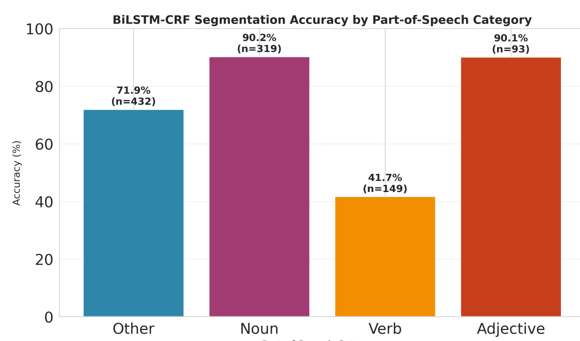

Figure 1. BiLSTM-CRF morphological segmentation accuracy by part-of-speech category, showing substantial variation in boundary detection performance across linguistic categories.

4.2. Morphological Similarity Analysis

We evaluate how well each tokenization approach preserves morphological relationships by measuring cosine similarity between lemmas and their inflected forms from the Uni-

Morph dataset:

$$\text{sim}(w_{lemma}, w_{wordform}) = \frac{v_{w_{lemma}} \cdot v_{w_{wordform}}}{||v_{w_{lemma}}|| \cdot ||v_{w_{wordform}}||} \tag{10}$$

When complete words are absent from model vocabularies, morpheme and BPE models generate compositional vectors by averaging constituent subword embeddings, while word-level models cannot handle out-of-vocabulary cases. BPE achieves the highest average similarity (0.73), followed by morpheme-based approaches (0.62) and word-level models (0.57). This unexpected result challenges the assumption that linguistically-informed segmentation inherently outperforms statistical methods.

Approach	Similarity (±SD)
Word	0.528 ± 0.15
Morpheme	0.583 ± 0.20
BPE	0.752 ± 0.18

Table 1. Average morphological similarity scores with standard deviations

BPE's apparently superior morphological similarity performance (0.752) must be interpreted cautiously due to severe evaluation coverage limitations. While BPE achieves higher average similarity with lower standard deviation (0.18), indicating more consistent scores, this performance is based on only 28.6% of test cases compared to 94.3% for word models.

The more structured embedding space organization demonstrated by morpheme models in neighbor rank analysis, combined with their superior clustering quality metrics, suggests that when fairly evaluated on comparable test sets, linguistically-informed segmentation may indeed outperform statistical approaches. The frequency-based patterns captured by BPE appear to create tighter morphological clustering for a limited subset of evaluable cases, but this advantage may not generalize to comprehensive morphological processing tasks. This coverage bias represents a critical methodological limitation that necessitates restricted evaluation on mutually evaluable word sets to enable fair comparison between tokenization approaches.

4.3. Segmentation Strategy Comparison

We quantify alignment between morpheme and BPE segmentation boundaries using the Jaccard similarity coefficient (Manning & Schütze, 1999). For each word w, we define A as morpheme boundary positions and B as BPE boundary positions:

$$\text{agreement}(w) = \frac{|A \cap B|}{|A \cup B|} \tag{11}$$

This metric ranges from 0 (no shared boundaries) to 1 (perfect boundary alignment), enabling analysis of convergence between statistical and linguistic segmentation approaches. Analysis reveals fundamental divergence between the two approaches. Despite processing identical text, morpheme and BPE methods achieve only 14.4% average boundary agreement, with 63.6% of words showing zero agreement and merely 2.5% achieving perfect alignment.

Metric	Value	Count
Zero agreement	63.6%	636 words
Perfect agreement	2.5%	25 words
Average agreement	14.4%	
Morpheme tokens/word	1.99	
BPE tokens/word	3.75	

Table 2. Segmentation agreement and tokenization density

The predominance of zero-agreement cases (63.6%) demonstrates that statistical and linguistic approaches identify almost entirely different sets of meaningful units. This fundamental divergence, combined with BPE's nearly doubled tokenization density, indicates that frequency-based patterns and morphological boundaries represent complementary rather than competing approaches to identifying linguistic structure. The minimal convergence suggests that hybrid methods combining both perspectives may be necessary to capture the full spectrum of meaningful units in Kurdish morphology.

4.4. Embedding Space Organization Analysis

We analyze embedding space structure through two complementary approaches: similarity distribution patterns and neighbor rank similarity analysis.

4.4.1. Neighbor Rank Similarity Analysis

We examine how similarity decreases across ranked nearest neighbors by calculating average similarity at each neighbor rank (Figure 2). Steep dropoff indicates well-organized semantic clusters, while flat curves suggest poorly structured embedding spaces. Contrary to expectations, morpheme and word models show nearly identical dropoff patterns, while BPE demonstrates the *shallowest dropoff* (i.e., flattest curve). All models start at similar similarity levels (0.88-0.89) for rank 1 neighbors, but BPE maintains consistently higher similarities across all neighbor ranks, ending at 0.54 at rank 20 compared to 0.50 for morpheme and word models.

The neighbor rank analysis reveals a counterintuitive pat-

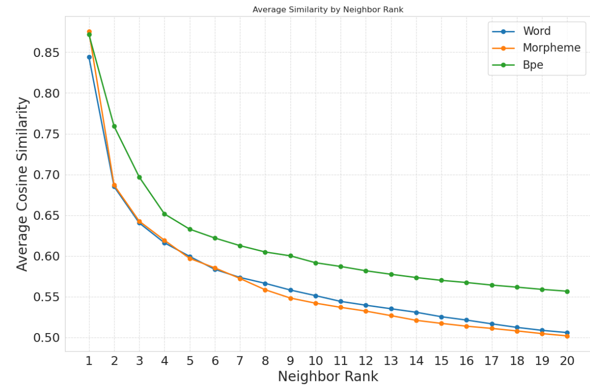

Figure 2. Average similarity by neighbor rank, showing how similarity decreases across ranked nearest neighbors for each tokenization approach. BPE maintains higher similarities across all ranks while morpheme and word models show steeper decay patterns.

Dropoff Measure	Word	Morpheme	BPE
Rank 1-5 Dropoff	31.2%	32.7%	30.3%
Rank 1-10 Dropoff	37.6%	38.8%	35.3%
Rank 1-20 Dropoff	43.1%	44.2%	39.1%

Table 3. Similarity dropoff rates at neighbor rank intervals

tern where BPE's morphological similarity scores coincide with less structured embedding space organization. BPE's shallow dropoff (39.1% rank 1-20) suggests more uniform semantic neighborhoods rather than the distinct clustering typically desired in embedding spaces. In contrast, morpheme and word models show steeper, more similar decay patterns (44.2% and 43.1% respectively), indicating better-defined semantic boundaries.

This finding challenges the interpretation of BPE's morphological similarity advantage, suggesting that higher similarity scores may reflect overly uniform embedding spaces rather than superior morphological understanding. The trade-off appears to favor consistency over discrimination: BPE creates embeddings where everything looks moderately similar to everything else, while morpheme-based approaches create more structured spaces with clearer semantic distinctions.

4.4.2. Allomorph Clustering Analysis

We analyze how different inflected forms of the same lemma cluster in embedding space by comparing intra-lemma distances (between forms of the same lemma) versus inter-lemma distances (between forms of different lemmas). This analysis reveals how well each tokenization approach groups morphologically related forms while maintaining separation between unrelated words.

The distance distribution plots (Figure 3) reveal distinct clustering patterns across tokenization approaches. All models achieve similar separation ratios (Word: 1.44, Morpheme: 1.31, BPE: 1.45), calculated as the ratio of average inter-lemma distances to average intra-lemma distances, indicating comparable ability to distinguish between different lemmas while clustering related forms. However, the distribution shapes differ substantially. BPE demonstrates the tightest intra-lemma clustering, with most related forms concentrated at very low distances (peak around 0.4-0.5). The morpheme model shows intermediate clustering behavior with a broader intra-lemma distribution, while the word model exhibits the most dispersed intra-lemma distances. All models maintain clear separation between intra-lemma and inter-lemma distances, with inter-lemma distributions consistently shifted toward higher distances (0.7-1.0).

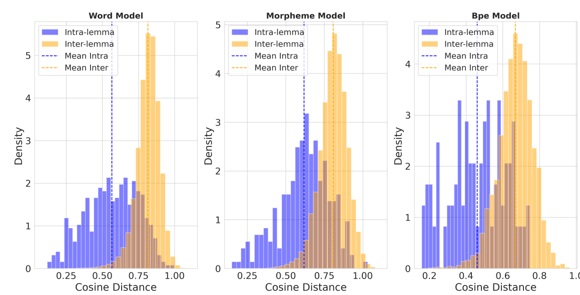

Figure 3. Distribution of cosine distances for intra-lemma (same lemma) versus inter-lemma (different lemmas) word pairs across tokenization approaches. Blue histograms show distances between inflected forms of the same lemma, while orange histograms show distances between forms of different lemmas.

Metric	Word	Morph	BPE
Separation Ratio	1.44	1.31	1.45
Cohesion	Moderate	Moderate	Highest
Pattern	Dispersed	Inter.	Concentrated

Table 4. Allomorph clustering characteristics

The allomorph clustering analysis provides crucial insight into BPE's apparent morphological similarity advantage. While BPE achieves the highest separation ratio and tightest intra-lemma clustering, this pattern may indicate overly uniform embeddings rather than superior morphological understanding. BPE's concentrated intra-lemma distribution suggests that morphologically related forms are clustered so tightly that fine-grained morphological distinctions may be lost.

This finding aligns with the neighbor rank analysis showing BPE's flatter similarity dropoff patterns. Together, these results suggest that BPE creates embedding spaces where morphologically related words are highly similar to each other, but this comes at the cost of reduced discriminative power and less structured semantic organization. The morpheme model's intermediate clustering behavior may represent a better balance between morphological coherence and semantic discrimination.

4.4.3. Similarity Distribution Analysis

We analyze the distribution of morphological similarity scores to understand each approach's evaluation coverage and bias patterns. Rather than examining only average performance, we investigate how frequently different similarity values occur across all lemma-wordform pairs to identify potential evaluation biases (Figure 4). BPE shows extreme concentration in high similarity ranges (55.4% of pairs achieve 0.8-1.0 similarity) with a sharp peak around 0.9. Word and morpheme models exhibit broader, more balanced distributions centered around 0.4-0.5.

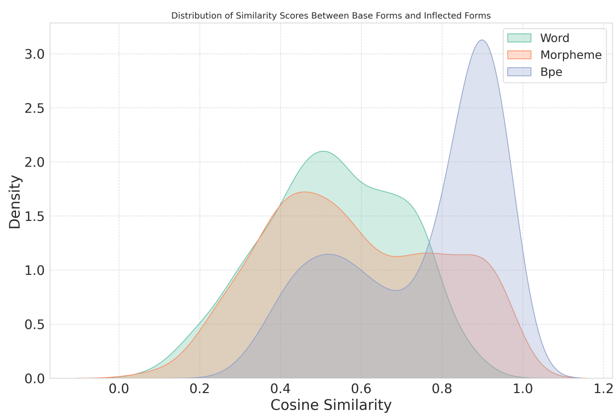

Figure 4. Similarity score distributions showing BPE's concentration in high-similarity ranges versus word and morpheme models' broader coverage.

Range	Word	Morph	BPE
0.0-0.4	23.4%	21.9%	3.9%
0.4-0.8	72.9%	57.0%	40.7%
0.8-1.0	3.7%	21.1%	55.4%

Table 5. Similarity score distribution by similarity range and tokenization strategy.

BPE's skewed distribution provides strong evidence for coverage bias. The sharp concentration at high similarities (55.4% vs 3.7% for word models) indicates BPE primarily evaluates morphological pairs with straightforward concatenative patterns where compositional vectors work well, such as regular stem-affix combinations. BPE fails to evaluate more challenging morphological relationships involving stem changes, irregular forms, or complex phonological processes that word and morpheme models can handle across the full complexity spectrum. This distribution pattern explains BPE's artificially inflated average similarity scores, as it essentially cherry-picks the morphological cases it can process successfully while being unable to evaluate the challenging cases that would reveal its limitations. The apparent superiority reflects selective evaluation of favorable cases rather than genuine morphological understanding.

4.5. Vocabulary Characteristics and Overlap Analysis

We analyze vocabulary sizes and overlap patterns across tokenization approaches to understand their fundamental differences in linguistic unit identification.

The three approaches produce dramatically different vocabulary characteristics. Word-level tokenization generates the largest vocabulary (260,922 tokens), morpheme-based segmentation produces an intermediate vocabulary (162,504 tokens), while BPE creates the smallest vocabulary (2,273 tokens), representing a 115-fold size difference.

Overlap analysis reveals striking patterns: morpheme and word models share substantial convergence (114,639 tokens, 70.5% of morpheme vocabulary), indicating many morphemes correspond to complete words, particularly for morphologically simple forms. In contrast, morpheme-BPE overlap is minimal (2,034 tokens, 1.3% of morpheme vocabulary), while BPE-word overlap appears high relative to BPE's small size (2,031 tokens, 89.4% of BPE vocabulary). This demonstrates that statistical and linguistic approaches identify almost entirely different sets of meaningful units.

Approach	Vocab Size	Coverage	Key Overlaps
Word-level	260,922	94.3%	–
Morpheme	162,504	68.7%	70.5% with Word model
BPE	2,273	28.6%	89.4% w/ Word model, 1.3% w/ Morph model

Table 6. Vocabulary and overlap patterns

4.6. Coverage Disparities and Evaluation Bias

The vocabulary differences documented above have profound implications for fair evaluation. Data coverage refers to the percentage of UniMorph lemma-wordform pairs each model can actually evaluate during testing (Figure 5). For a morphological similarity assessment to be computed, both the lemma and its inflected form must have available embeddings, either through direct vocabulary lookup or through compositional vector construction from subword components.

This coverage analysis reveals severe disparities: BPE can evaluate only 28.6% of UniMorph test cases, while morpheme models handle 68.7% and word-level models achieve 94.3% coverage. These differences introduce systematic evaluation bias, as models are essentially being tested on different subsets of morphological complexity.

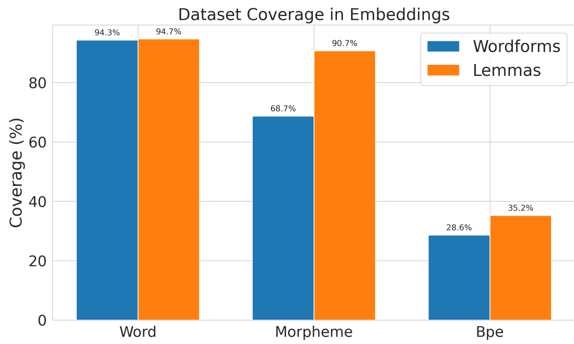

Figure 5. Dataset coverage in embeddings showing dramatic differences in evaluation coverage across tokenization approaches

BPE's apparent superior performance (0.752 average similarity) may reflect evaluation on a highly selective subset where compositional vector construction succeeds, primarily cases with straightforward concatenative morphology. Meanwhile, word and morpheme models are evaluated across substantially larger and more representative portions of the morphological complexity spectrum. This coverage bias fundamentally compromises the validity of direct performance comparisons and suggests that BPE's morphological similarity advantage may be artifactual rather than substantive.

5. Conclusion and Future Work

5.1. Summary of Key Findings

Our evaluation of Kurdish tokenization strategies reveals nuanced patterns that challenge simple assumptions about morphological segmentation effectiveness. While BPE initially appears to outperform morpheme-based approaches (0.752 vs 0.583 average morphological similarity), this advantage is fundamentally compromised by severe evaluation coverage limitations. BPE's evaluation coverage is severely limited (28.6% of test cases) compared to morpheme models (68.7%) and word models (94.3%), suggesting its superior performance reflects selection bias toward favorable cases rather than genuine morphological understanding. The segmentation agreement analysis confirms fundamental divergence between approaches, with only 14.4% average boundary agreement and 63.6% zero-agreement cases. Despite processing identical text, morpheme and BPE methods identify almost entirely different meaningful units (1.3% vocabulary overlap), with BPE producing nearly twice the tokenization density (3.75 vs 1.99 tokens per word). Critically, multiple embedding quality measures beyond morphological similarity reveal BPE's limitations. The neighbor rank analysis shows BPE creates less structured embedding spaces with flatter similarity dropoff patterns (39.1% vs 44.2% for morphemes), while clustering quality metrics favor word and morpheme models. These findings suggest BPE's apparent morphological advantage comes at the cost of overall semantic organization and discriminative power.

5.2. Broader Implications

Our results highlight the complexity of evaluating tokenization approaches for morphologically rich languages. The coverage bias phenomenon demonstrates that high similarity scores can be misleading when based on selective evaluation subsets. This finding has broader implications for low-resource language processing, where compositional vector approaches may systematically exclude challenging morphological relationships.

The minimal agreement between statistical and linguistic segmentation approaches (14.4%) suggests these methods capture complementary rather than competing aspects of linguistic structure. Frequency-based patterns and morphological boundaries may represent different but equally valid perspectives on meaningful unit identification, necessitating hybrid approaches for comprehensive coverage.

The BiLSTM-CRF segmentation results reveal substantial variation across part-of-speech categories (90.2% for nouns vs 41.7% for verbs), creating natural experiments for understanding how segmentation quality propagates to embedding performance. This POS-specific variation provides insights into the linguistic complexity that tokenization systems must address.

5.3. Future Directions

Future work should focus on developing fair comparison methodologies across tokenization approaches with different coverage characteristics, including restricted evaluation frameworks and coverage-robust similarity metrics. Given the complementary nature of statistical and linguistic approaches, exploring morphologically-constrained BPE algorithms and multi-level frameworks that combine both insights represents a promising direction. The poor verb segmentation performance (41.7%) indicates need for specialized Kurdish morphological processing through verb-specific models, phonological change rules, and complex affixation pattern handling. Practical validation through downstream Kurdish NLP tasks would complement intrinsic measures, while investigating tokenization impacts on transformer-based models offers opportunities for Kurdish-specific language model development. Future advances will emerge from hybrid methods combining different tokenization paradigms' strengths while addressing individual limitations.

6. Limitations

The BiLSTM-CRF morphological segmentation shows poor performance on Kurdish verbs (41.7% accuracy), potentially undermining morpheme-based approaches' effectiveness. However, this may reflect our training methodology rather than inherent constraints. The BiLSTM-CRF model was trained using random sampling with bootstrap annotation, without explicit control over part-of-speech distribution. Given the model's strong performance on nouns (90.2%) and adjectives (90.1%), the poor verb segmentation may reflect insufficient exposure to verbal morphological patterns during training rather than fundamental model limitations. A more balanced training approach with deliberate inclusion of diverse verb forms might significantly improve segmentation performance and, consequently, morpheme-based embedding quality. The evaluation employs intrinsic similarity measures, which provide controlled assessment of morphological relationship preservation but would benefit from complementary downstream task validation to demonstrate practical applicability. However, comprehensive downstream evaluation is constrained by the limited availability of Kurdish NLP datasets and task-specific resources, a common challenge in low-resource language research.

References

Ahmadi, S. A tokenization system for the Kurdish language. In Zampieri, M., Nakov, P., Ljubešić, N., Tiedemann, J., and Scherrer, Y. (eds.), *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 114–127, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL). URL https://aclanthology.org/2020.vardial-1.11/.

Ahmadi, S. Hunspell for sorani kurdish spell checking and morphological analysis. *ArXiv*, abs/2109.06374, 2021. URL https://api.semanticscholar.org/CorpusID:237503500.

Ahmadi, S. and Hassani, H. Towards finite-state morphology of kurdish. *ArXiv*, abs/2005.10652, 2020. URL https://api.semanticscholar.org/CorpusID:218763544.

Ahmadi, S. and Wurm, M. L. Tokenization for low-resource languages: A case study on Kurdish. *arXiv preprint arXiv:1909.07739*, 2019.

Ahmadi, S., Hassani, H., and McCrae, J. P. Towards electronic lexicography for the Kurdish language. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 9–18, 2020.

Bostrom, K. and Durrett, G. Byte pair encoding is suboptimal for language model pretraining. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4617–4624, 2020.

Cotterell, R. and Schütze, H. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1660, 2016.

Creutz, M. and Lagus, K. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Helsinki University of Technology, 2005.

Erdmann, A. and Habash, N. Complementary strategies for low resourced morphological modeling. In Kuebler, S. and Nicolai, G. (eds.), *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 54–65, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5806. URL https://aclanthology.org/W18-5806/.

Esmaili, K. S. and Salavati, S. Sorani Kurdish versus Kurmanji Kurdish: An empirical comparison. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 300–305, 2013.

Gerz, D., Vulic, I., Ponti, E. M., Naradowsky, J., Reichart, R., and Korhonen, A. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 316–327, 2018.

Hassani, H. BLARK for multi-dialect languages: towards the kurdish BLARK. *Language Resources and Evaluation*, 52(2):625–644, 2018.

Huang, Z., Xu, W., and Yu, K. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991, 2015. URL https://api.semanticscholar.org/CorpusID:12740621.

Kudo, T. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, 2018.

Kudo, T. and Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018.

Lafferty, J., McCallum, A., and Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning*, pp. 282–289, 2001.

Luong, T., Socher, R., and Manning, C. D. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113, 2013.

Mahmudi, A. and Veisi, H. Automated grapheme-to-phoneme conversion for central kurdish based on optimality theory. *Computer Speech Language*, 70:101222, 2021. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2021.101222. URL https://www.sciencedirect.com/science/article/pii/S0885230821000292.

Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Mielke, S. J., Cotterell, R., Gorman, K., Roark, B., and Eisner, J. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4975–4989, 2019.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., and Schwartz, L. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276, 2020. URL https://doi.org/10.1162/tacl_a_00365.

Pimentel, T., Ryskina, M., Mielke, S. J., Wu, S., Chodroff, E., Leonard, B., Nicolai, G., Ghanggo Ate, Y., Khalifa, S., Habash, N., El-Khaissi, C., Goldman, O., Gasser, M., Lane, W., Coler, M., Oncevay, A., Montoya Samame, J. R., Silva Villegas, G. C., Ek, A., Bernardy, J.-P., Shcherbakov, A., Bayyr-ool, A., Sheifer, K., Ganieva, S., Plugaryov, M., Klyachko, E., Salehi, A., Krizhanovsky, A., Krizhanovsky, N., Vania, C., Ivanova, S., Salchak, A., Straughn, C., Liu, Z., Washington, J. N., Ataman, D., Kieraś, W., Woliński, M., Suhardijanto, T., Stoehr, N., Nuriah, Z., Ratan, S., Tyers, F. M., Ponti, E. M., Aiton, G., Hatcher, R. J., Prud'hommeaux, E., Kumar, R., Hulden, M., Barta, B., Lakatos, D., Szolnok, G., Ács, J., Raj, M., Yarowsky, D., Cotterell, R., Ambridge, B., and Vylomova, E. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 229–259, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigmorphon-1.25. URL https://aclanthology.org/2021.sigmorphon-1.25.

Řehůřek, R. and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA.

Ruder, S., Vulić, I., and Søgaard, A. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65 (1):569–630, May 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL https://doi.org/10.1613/jair.1.11640.

Schuster, M. and Nakajima, K. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152. IEEE, 2012.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, 2016.

Sheykh Esmaili, K. and Salavati, S. Sorani Kurdish versus Kurmanji Kurdish: An empirical comparison. In Schuetze, H., Fung, P., and Poesio, M. (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 300–305, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2054/.

Smit, P., Virpioja, S., Grönroos, S.-A., and Kurimo, M. Morfessor 2.0: Toolkit for statistical morphological segmentation. In Wintner, S., Tadić, M., and Babych, B. (eds.), *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 21–24, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-2006. URL https://aclanthology.org/E14-2006/.

Thackston, W. M. *Sorani Kurdish: A Reference Grammar with Selected Readings*. Harvard University Press, Cambridge, MA, 2006.

Veisi, H., MohammadAmini, M., and Hosseini, H. Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193, 02 2019. ISSN 2055-7671. doi: 10.1093/llc/fqy074. URL https://doi.org/10.1093/llc/fqy074.

Appendix

Corpus Preprocessing and Normalization

The Kurdish corpus required extensive preprocessing to address orthographic inconsistencies and dialectal variations inherent in Kurdish text collections. This section includes more details about the normalization procedures applied to ensure consistent tokenization and analysis.

Character Repetition Normalization: User-generated content frequently contained repeated characters for emphasis (e.g., سڵاووووو for emphasis). We implemented a normalization rule that reduced any character sequence repeated more than three times to exactly three repetitions, thus سڵاووووو became سڵاووو. This approach balanced noise reduction with preservation of linguistic nuance, as triple repetition often carries semantic meaning in Kurdish.

Script and Character Standardization: Sorani Kurdish exhibits pervasive orthographic inconsistencies in some characters, particularly with characters ە and ه, which represent the same phoneme but are used differently based on context, keyboard layout, and individual preference. These variations often involve different Unicode code points, creating artificial vocabulary inflation. We developed a custom character replacement system that mapped these variants to canonical forms, ensuring consistent representation throughout the corpus.

Zero-Width Non-Joiner (ZWNJ) Handling: Arabic-based script languages like Kurdish frequently contain zero-width non-joiner characters that alter character joining behavior without visible effect, disrupting consistent tokenization. Different keyboard layouts encode ZWNJ differently. Some through dedicated keys, others through key combinations—leading to inconsistent usage patterns. Users sometimes substitute full spaces or omit the character entirely, creating tokenization ambiguity. We implemented corpus-wide ZWNJ regularization through text-level normalization routines that either removed or standardized these characters based on contextual appropriateness.

Quality Filtering: We applied AsoSoft's text normalization method from their Python library in initial preprocessing stages. Subsequently, sentences shorter than 5 characters were removed to eliminate fragments and malformed entries. Additionally, lines lacking valid Kurdish characters (identified using Unicode ranges for Kurdish script) were filtered out to ensure corpus linguistic consistency.

Dialectal Filtering: The corpus contained multi-dialectal interference from Kurmanji Kurdish and Persian sources. We developed character profile analysis to identify sentences deviating from Sorani Kurdish orthographic norms, filtering out content that exhibited non-Sorani characteristics. This process helped maintain dialectal consistency while preserving corpus size.

Format Standardization: Final preprocessing involved segmenting documents into sentence-per-line and word-per-line formats to support various downstream tasks including morphological segmentation training and embedding model development. This dual-format approach enabled flexible corpus utilization across different experimental requirements.

Implementation Details

All experiments were conducted using Python 3.9.6 with PyTorch 1.12.0 for neural model implementation. The BiLSTM-CRF morphological segmenter was trained using the torchcrf library (1.1.0) with Adam optimizer and learning rate scheduling. word2vec models were trained using Gensim 4.2.0 with identical hyperparameters across tokenization approaches to ensure fair comparison.

BPE tokenization employed the Hugging Face tokenizers library (0.15.2) with vocabulary size set to 2,280 tokens based on empirical morpheme count analysis. The tokenizer was trained on the full corpus before applying to sample data for embedding training. Morphological segmentation used our BiLSTM-CRF model with 3-layer bidirectional LSTM (hidden size 128) and CRF output layer.

Experimental Configuration

word2vec training employed skip-gram architecture with negative sampling (5 negative samples), vector dimensions of 150, minimum count threshold of 5, and 10 training epochs. Window sizes were adjusted proportionally based on tokenization density: word-level (5), morpheme-based (10), and BPE (19). All models used identical random seeds (42) for reproducibility.

The evaluation corpus comprised 1.5 million sentences sampled from a larger Kurdish text collection (Veisi et al., 2019). UniMorph evaluation used 996 Kurdish lemma-wordform pairs with morphological feature annotations. Coverage analysis and similarity computations employed cosine similarity with compositional vector generation for out-of-vocabulary terms through subword averaging.

Computational Resources

Experiments were conducted on systems with NVIDIA RTX 3080 GPUs and 32GB RAM. BiLSTM-CRF training required less than an hour, while Word2Vec training ranged from 4 minutes (word-level) to 25 minutes (BPE) depending on vocabulary size and tokenization complexity.

All code and trained models will be available to facilitate reproduction and extension of these results.