The effect of model capacity and script diversity on subword tokenization for Soranî Kurdish

Ali Salehi **Cassandra L. Jacobs**
Department of Linguistics
University at Buffalo
asalehi;cxjacobs@buffalo.edu

Abstract

Tokenization and morphological segmentation continue to pose challenges for text processing and studies of human language. Here, we focus on written Soranî Kurdish, which uses a modified script based on Persian and Arabic, and its transliterations into the Kurdish Latin script. Importantly, Perso-Arabic and Latin-based writing systems demonstrate different statistical and structural properties, which may have significant effects on subword vocabulary learning. This has major consequences for frequency- or probability-based models of morphological induction. We explore the possibility that jointly training subword vocabularies using a source script along with its transliteration would improve morphological segmentation, subword tokenization, and whether gains are observed for one system over others. We find that joint training has a similar effect to increasing vocabulary size, while keeping subwords shorter in length, which produces higher-quality subwords that map onto morphemes.

1 Introduction

Different scripts for the same language may convey different linguistic and structural properties, such as phonological transparency, word boundaries (e.g., whitespace), morpheme boundaries, serial position within a word, or present different orthotactic and spelling constraints. In this work, we examine the relationship between script variation, morphological acquisition, and subword vocabulary construction. Obtaining high-quality morphological annotations is critical for linguistic analysis, so unsupervised methods for learning morpheme-like representations are often an acceptable compromise. Here, we explore the usefulness of subword vocabulary training for morphological segmentation of written Soranî Kurdish, a central dialect of Kurdish spoken mainly in Iran and Iraq. Soranî is morphologically complex but relatively underresourced, with few large annotated corpora (Veisi et al., 2019; Malmasi, 2016; Goldhahn et al., 2012; Ahmadi, 2020a; Mahmudi and Veisi, 2021), and none with adequate morphological glosses or segmentations for downstream language model development (Alkaoud and Syed, 2020; Banerjee and Bhattacharyya, 2018) or linguistic analysis.

Soranî has some unlabeled raw text corpora, which opens the possibility to leverage the statistical properties of the text for unsupervised subword-based vocabulary induction. The existence of multiple writing systems for Kurdish languages additionally presents a challenge for NLP systems, and jointly training subword tokenization models may be advantageous for Central Kurdish NLP in general. We thus ask whether training a subword vocabulary on multiple scripts can induce adequate morphological segmentations and compare such systems against models trained solely on single scripts of equivalent or larger sizes.

2 Soranî morphology and script variation

Our manipulation leverages script variability in the written Kurdish dialects. Kurdish dialects have been written with diverse writing systems including Arabic, Latin, Yekgirtú (unified), Cyrillic and Armenian scripts. There is no unified orthography for Kurdish despite previous efforts (Ahmadi et al., 2020). This variation presents an intriguing opportunity to explore the impact of input diversity on the learning of subword vocabulary. Furthermore, demonstrating the potential usefulness of joint training on multiple scripts could produce higher-quality multi-dialect Transformer language models (Kanjirangat et al., 2023).

The Soranî writing system used for Central and Southern Kurdish is written in a modified Perso-Arabic script and has an alphabetic structure with a high degree of phonological transparency, relative to Arabic and Persian scripts (Chyet and Schwartz, 2003; Ahmadi, 2020b). The Latin-based Hawar al-

Proceedings of the 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 51–56

phabet, used by Northern Kurdish dialects, shares this transparency, making it feasible to transliterate Soranî script into a Latin-based one (Mahmudi and Veisi, 2021). The Latin script has two allographs per segment (e.g., H/h), which mostly encode sentence position, but the Perso-Arabic script has three for word-initial, -medial, and -final positions (e.g., the phoneme /h/ is represented by 'هـ' word initially, by 'ـهـ' word medially and by 'ه' or 'ـه' word finally.

3 Subword tokenization

We explore multiple tokenization models for Soranî Kurdish, with a primary focus on Byte Pair Encoding (BPE; Sennrich et al., 2016a) and Unigram tokenization (Kudo, 2018). We assessed these models' performance with respect to morphological and phonological structure and critically assess claims of better morphological induction by Unigram relative to BPE (Bostrom and Durrett, 2020). Both BPE and Unigram LM have probabilistic components based on frequency and vocabulary likelihood, respectively. These models were selected for their ability to handle diverse linguistic data and to learn meaningful linguistic units from large datasets without extensive annotated resources. The focus on these tokenization models that are used in modern neural methods is to align any prospective tokenization system with current trends, enabling scalability and robustness across different datasets and providing an exploration of their adaptability to the morphological richness of Kurdish in both Latin and Arabic scripts.

Byte-Pair Encoding. BPE is a simple and common subword tokenization algorithm (Gage, 1994; Sennrich et al., 2016a) that grows a vocabulary from individual characters into more complex subwords by merging the most frequent co-occurring character sequence, up to a specified number of merges. Given the frequency-based merging process of BPE, it is plausible that manipulating the relative frequency of subwords by training multiple scripts will influence the resulting vocabularies. We further hypothesize that the different character frequency distributions of Latin and Perso-Arabic scripts may help BPE to learn subwords that better align with morphological boundaries and better capture the tendency of non-stem morphemes in Kurdish to be short.

Unigram tokenization. This subword method iteratively splits words into subwords by optimizing the likelihood of the training data, which providing a probabilistic approach to subword segmentation that may capture more nuanced linguistic patterns compared to BPE's frequency-based merging strategy (Kudo, 2018). Unigram tokenization has been argued to produce better morphological segmentations than algorithms like BPE or WordPiece (Bostrom and Durrett, 2020). We expect Unigram tokenization to potentially provide more comprehensive coverage of Soranî morphology compared to BPE, due to the likelihood objective of Unigram.

4 Experiments

We used the huggingface tokenizer package for BPE (Sennrich et al., 2016b) and Unigram tokenization (Kudo and Richardson, 2018). For our experiments, we used the normalized version of the Asosoft corpus (Veisi et al., 2019) consisting of 188 million word tokens and 4.66 million word types. The corpus was chosen for its comprehensive coverage of the Soranî dialect. The corpus includes 58,000 documents from textbooks and magazines and 400,000 documents from web crawls. We removed newline characters, repeated characters (Rajadesingan et al., 2015), and redundant whitespace before subword training. We tested separate models for Latin and Arabic scripts, each with a 5k vocabulary size. Additionally, we constructed a joint Arabic-Latin script corpus for data augmentation and further constrained model size based on script-specific vocabulary sizes derived from this joint corpus.

Vocabulary size. We explored various vocabulary sizes within a range of 1,000 to 10,000 subwords to identify the optimal balance between granularity and generalization. For BPE, we found that a larger vocabulary size of 5,000 subwords provided the best results, and so we use this size in all our experiments. This size qualitatively offered a good trade-off between identifying roots and affixes versus learning morphologically complex words, capturing the morphological complexity of Soranî in both Latin and Arabic scripts. Across all of our measures, 5,000 subwords each for the Latin and Arabic scripts led to the highest performance.

Transliteration. The Latin-based script exhibits a one-to-one correspondence between phonemes and alphabet letters (Esmaili et al., 2013) that can be deterministically transliterated from the Arabic script using Asosoft (Mahmudi and Veisi, 2021). In addition to changing character frequencies caused

Model	Vocab Size	Script	Avg. Len.	Tokenization Agreement (%)	Syllabification (%)
BPE - Small	2514	Latin	2.94	75.29	
	2446	Arabic	2.87		4.87
BPE - Large	5000	Latin	3.46	**79.67**	
	5000	Arabic	3.38		12.64
BPE - Joint	2514	Latin	3.04	77.08	
	2446	Arabic	2.98		**26.70**
Unigram - Large	5000	Latin	3.33	74.28	
	5000	Arabic	3.24		11.73
Unigram - Joint	3892	Latin	3.09	76.72	
	3647	Arabic	3.01		25.77

Table 1: Comparison of tokenization models for Soranî Kurdish in Latin and Arabic scripts.

by multiple allographs, transliteration into the Latin script introduces the letter "i" for the schwa, which is not encoded in the Perso-Arabic script. The relative transparency of the Latin script may produce more accurate segmentations than the Arabic script.

Data "augmentation." We define a joint tokenizer as a tokenization model trained on text data from multiple scripts simultaneously. For Kurdish languages, which can be written in both Latin and Arabic scripts, a joint tokenizer aims to create a unified subword vocabulary that can effectively tokenize text for multiple dialects. This approach combines text data from both scripts for a balanced training set, which the BPE and Unigram model then uses to develop a script-agnostic tokenization strategy based on subword frequency. This effectively doubles the training data set size and may alter the relative frequencies of subwords in the data. We measure the different tokenizers' precision against verified morphological segmentations of Soranî Kurdish, along with segmentation accuracy. We hypothesize changes in subword tokenization following from the fact that the two scripts have slightly different orthotactics (see Section 2). Transliteration is hypothesized to enhance subword vocabulary training by increasing the number of data points under consideration (Shazal et al., 2020; Biadgligne and Smaili, 2023).

5 Results

5.1 Subword vocabularies

We first characterize the subword vocabularies and their behavior for words in the training corpus. Our analysis includes the token match rate between Latin and Arabic scripts, average token length, syllable-token correspondence, and token-morpheme match rate to assess the effectiveness of subword tokenization models in capturing the linguistic structure of Soranî Kurdish (Table 1).

Token length. The average length of the tokens reveals the granularity of the subword segmentation, with shorter lengths indicating finer segmentation. Unigram model tends to produce longer subwords, indicating differences in the granularity of tokenization caused by the split procedure. The BPE models produce shorter subwords, which follows given the merge-based training procedure, and this is especially true for small, single-script models.

Tokenization consistency. The percentage of tokenizations that are at the same boundaries across both Arabic and Latin scripts (Token % in Table 1) highlights the models' ability to maintain consistency across different writing systems, which is crucial for script-agnostic NLP applications. It is meant to measure the consistency of tokenization by comparing boundary positions in both scripts, quantifying the percentage of boundaries that coincide. The larger independently-trained BPE models achieve the highest token match rate at 79.67%, suggesting that similar types of merges are occurring for both scripts.

Syllabification. Syllable-token correspondence measures the alignment of tokens with the syllabic structure of the language. The highest percentage of matching occurs with BPE trained jointly.

5.2 Morphological coverage

We assess the quality of the subword vocabularies by computing the overlap between the tokens generated by BPE and the morphemes of the words, as well as the proportion of token strings that cor-

Setting	Model	Mean tokens in morpheme set	Mean morphemes in vocabulary	Morpheme Coverage %	Segmentation Accuracy %
Latin Script					
Joint training	BPE	0.349	0.345	43	26
	Unigram	*0.428*	*0.423*	47	*34*
2514 subwords	BPE	0.336	0.333	41	25
5000 subwords	BPE	0.379	0.370	*50*	29
	Unigram	**0.440**	**0.435**	**51**	**36**
Arabic Script					
Joint training	BPE	0.368	0.361	44	28
	Unigram	*0.484*	*0.479*	49	*40*
2446 subwords	BPE	0.353	0.349	41	26
5000 subwords	BPE	0.402	0.390	*52*	32
	Unigram	**0.496**	**0.485**	**54**	**43**

Table 2: Performance metrics of tokenization models for Soranî Kurdish.

respond to morphemes and the proportion of morphemes that are present in the subword vocabulary.

To create the test set for evaluating the tokenization models, we selected words from the corpus that represent a variety of linguistic phenomena in Soranî. This included words with ezafe constructions, compounds, preverbal constructions, and words that incorporate prepositions. We also chose words that contain a half space or Zero Width Non-Joiner (ZWNJ) to assess the models' ability to handle this aspect of the script. To evaluate the models' performance in capturing the morphological structure of Soranî, 1500 words were manually tokenized to accurately segment the morphemes. Table 3 illustrates the efficacy of different tokenization models in segmenting Soranî words into their respective morphemes. We compare Unimorph and BPE with different vocabulary sizes, across a selection of words. The comparison focuses on how each model tokenizes the words and aligns these tokens with the linguistically motivated morpheme boundaries. For instance, the word "destîpêkird" is tokenized differently by Unimorph and BPE, reflecting each model's approach to parsing the underlying morphological structure of the language.

Combining two scripts has a small but positive effect on tokenization quality in terms of morphological accuracy for BPE, relative to the small single-script models. When BPE is trained to a larger subword vocabulary for either script, it performs slightly better in terms of morphological coverage compared to other models, including the joint BPE model. This highlights the potential trade-offs between vocabulary size and script coverage in subword tokenization for Soranî Kurdish. However, Unigram tokenization consistently outperforms on all measures of morphological structure, as seen in prior work (Bostrom and Durrett, 2020). We summarize these comparisons in Table 2.

6 Future work

A specific subset of words containing the Zero-width non-joiner (ZWNJ) was deliberately isolated to assess tokenization performance of the Unigram tokenizer with 5000 subwords, particularly within the Perso-Arabic script. The presence of ZWNJ, which can act as a morphological delimiter in word or appear after the letter 'ه' /a/ by pressing the E key on the keyboard, helps in achieving more accurate segmentation outcomes. For instance, in the word *tokmetir* 'تۆکمەتر' 'stronger', the word *tokme* 'تۆکمە' is separated from the comparative morpheme *tir* 'تر' by a ZWNJ which gets tokenized as a two actual morphemes by the Unigram tokenizer. This structural feature can provide clues to tokenization models, enabling more precise identification and segmentation of morphemes and a higher granularity in morpheme segmentation compared to the Latin script. Such findings underscore the importance of leveraging script-specific orthographic cues to improve tokenization models for

Tokenizer	vocab	word	Latin Tokens		Morphemes
Unimoph	5k	destîpêkird	[‘destîpêkird’]	[دەستپێکرد]	dest-î-pê-kird
Unimorph	5k	meseley	[’meseley’]	[مەسەلەی]	mesele-y
Unimorph	5k	pîlangêřîyekan	[‘pîlan’, ‘gêř’, ‘îyekan’]	[پیلان، گێڕ، سەکان]	pîlan-gêřî-yek-an
BPE	5k	destîpêkird	[‘destî’, ‘pêkird’]	[دەستی، پێکرد]	dest-î-pê-kird
BPE	5k	lebexda	[‘lebexda’]	[لەبەغدا]	le-bexda
BPE	2514/2446	damezrawekanî	[‘damez’, ‘rawe’, ‘kanî’]	[دامەز، راو، ەکانی]	damezraw-ekan-î
BPE	2514/2446	lelayekewe	[’lelay’, ’ek’, ’ewe’]	[لەلای ، ەک ، ەوە]	yekêtî-ye
BPE	2514/2446	gořanêk	[‘gořan’, ‘êk’]	[گۆڕان، تک]	gořan-êk

Table 3: Token and morpheme segmentation examples across Unimorph and BPE tokenizers

under-resourced language contexts.

While recognizing the contributions of traditional unsupervised segmenters such as Morfessor (Creutz and Lagus, 2005), Adaptor Grammars (Johnson et al., 2006) and DPSeg (Dirichlet Process-based Segmenter) (Goldwater et al., 2005) in morphological analysis, this research primarily explores the application these subword tokenizers that are used in modern neural methods. We will extend this comparison to include these traditional segmenters, particularly focusing on their unigram versions which share similarities with the Unigram model used in this study. For future work, we wish to explore the effects of using smaller training datasets with less bias in frequency distribution, build tokenization models based on vocabularies rather than corpora, and train greedy contextual decoding tokenizers (e.g., Uzan et al., 2024).

7 Conclusion

In this study, we have explored the capacity of different tokenization models to segment Soranî Kurdish text into morphologically well-formed subwords. Our findings highlight the differentical effects of pruning and merging on the inductive biases of these models, shedding light on their ability to capture morphological structures. We find that Unigram tokenization leads to the highest quality off-the-shelf morpheme segmentation and find that data augmentation is a less effective strategy than increased vocabulary size in a monoscript context. This research will contribute to the development of more effective NLP tools for low-resource languages with smaller sources and only vocabulary lists, with a focus on morphologically and phonologically motivated analyses.

References

Sina Ahmadi. 2020a. KLPT – Kurdish language processing toolkit. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 72–84, Online. Association for Computational Linguistics.

Sina Ahmadi. 2020b. A tokenization system for the Kurdish language. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–127, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Sina Ahmadi, Hossein Hassani, and Kamaladdin Abedi. 2020. A corpus of the Sorani Kurdish folkloric lyrics. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 330–335, Marseille, France. European Language Resources association.

Mohamed Alkaoud and Mairaj Syed. 2020. On the importance of tokenization in Arabic embedding models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 119–129, Barcelona, Spain (Online). Association for Computational Linguistics.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Yohannes Biadgligne and Kamel Smaili. 2023. Baseline transliteration corpus for improved english-amharic machine translation. *Informatica*, 47(6).

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Michael L Chyet and Martin Schwartz. 2003. *Kurdish-English Dictionary*. Yale University Press.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. *Helsinki University of Technology*.

Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi. 2013. Building a test collection for sorani kurdish. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Leipzig corpora collection. Available online at https://corpora.uni-leipzig.de/.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2005. Interpolating between types and tokens by estimating power-law generators. In *Advances in neural information processing systems*.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in neural information processing systems*.

Vani Kanjirangat, Tanja Samardžić, Ljiljana Dolamic, and Fabio Rinaldi. 2023. Optimizing the size of subword vocabularies in dialect classification. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 14–30, Dubrovnik, Croatia. Association for Computational Linguistics.

Taku Kudo. 2018. 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)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. 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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Aso Mahmudi and Hadi Veisi. 2021. Automated grapheme-to-phoneme conversion for central kurdish based on optimality theory. *Computer Speech Language*, 70:101222.

Shervin Malmasi. 2016. Subdialectal differences in Sorani Kurdish. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 89–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 97–106, New York, NY, USA. Association for Computing Machinery.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for Arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 167–177, Barcelona, Spain (Online). Association for Computational Linguistics.

Omri Uzan, Craig W. Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is all you need: An evaluation of tokenizer inference methods.

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2019. Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.

8 Appendix

The python version used in this paper is 3.9.6. The Hugging Face tokenizer library version 0.15.2 is used for training BPE and Unigram models. Sentencepiece is trained using version 0.2.0.