The Good, The Bad, and The Features: Comparing Affix-Based and BPE-Based Feature Design for Morphological Tagging

Gustaf Gren
`gustaf.gren@ling.su.se`

April 9, 2025

Abstract

This study pits two classifiers — multinomial logistic regression and a perceptron — against each other for morphological tagging, each armed with a different set of features: affixes and byte pair encoding (BPE). While neither model has the sleight of hand to shift the outcome, BPE shows to either perform the same, or worse, across languages and metrics.

1 Background and Motivation

Morphology studies the structure and formation of words, including roots, prefixes, suffixes, and inflections. Computational morphology uses computational methods to analyze these aspects, such as *Morfessor* (Smit et al. 2014) which automatically segments text. Morphological tagging automatically labels words with their morphological tags, but the impact of model choice on performance remains underexplored. Haley, Ponti, and Goldwater (2023) compared regression and perceptrons for inflection/derivation classification, both reaching 90% accuracy.

Recent research suggests BPE subwords align with morphological features. Gutierrez-Vasques, Bentz, and Samardi (2023) linked early BPE merges to morphological typology, while Saleva and Lignos (2021) found no significant difference between BPE and morphology-based segmentation in low-resource translation.

This article examines how model choice (multinomial logistic regression/perceptron) and feature choice (affix-based vs. BPE-based) impact morphological tagging performance on three languages from the UniMorph dataset (Batsuren et al. 2022): Welsh (fusional), Chichimeca Jonaz (tonal), and Basque (agglutinative).

2 Methodology

The open source Python library `TiNLP` was used for these experiments, with reproducible scripts here: `https://github.com/skogsgren/tinlp`

2.1 Data

UniMorph data has three columns: lemma, inflected word, and part-of-speech & morphological tag. This article only uses the second and third columns, predicting only the morphological tag (e.g., `DEF;NOM;SG`), with `N` riding shotgun as an extra feature. Each data was randomly sampled to a 80/20 train/test split.

2.2 Models & Evaluation

A perceptron is a simple neural network using a linear classifier on input features. We use two feature functions: one uses affixes ($1 \rightarrow n$ characters from a word's start/end), and the other subwords from a low-vocab BPE model (see appendix A for pseudocode). The multinomial logistic regression model architecture follows Jurafsky and Martin (2025, Ch.5).

Models are trained for 3 epochs per language with affix length 5 and BPE vocab size 500 respectively. Evaluation includes accuracy and macro F1-score, followed by paired bootstrap tests (Tibshirani and Efron 1993) for each model and metric, with 5000 total samples each.

3 Results

Except for Basque, model choice did not significantly affect morphological tagging performance on either metric or feature function (see tables 3-6). Since the model choice was not statistically significant, the perceptron was selected for comparison between affix-based and BPE-based features due to its faster training and inference. Results show that the perceptron trained on affix-based features outperforms the model trained on low-vocab BPE-based features in most languages, except Welsh, where both models had identical performance. Full metrics are presented in table 1-2. All tables are rounded to two decimals.

Language	Affix Accuracy	BPE Accuracy	δ	p	95% CI
Welsh	0.73	0.73	0.00	0.642	$[-0.023 \quad 0.017]$
Chichimeca	0.52	0.43	0.09	<0.001	$[0.066 \quad 0.105]$
Basque	0.30	0.12	0.17	<0.001	$[0.151 \quad 0.191]$

Table 1: Accuracy morphological tagging performance with paired bootstrap test (H_0: no difference between affix and BPE-based features)

Language	Affix F1	BPE F1	δ	p	95% CI
Welsh	0.71	0.71	0.00	0.369	$[-0.012 \quad 0.022]$
Chichimeca	0.50	0.40	0.10	<0.001	$[0.085 \quad 0.12]$
Basque	0.22	0.08	0.14	<0.001	$[0.118 \quad 0.146]$

Table 2: F1 morphological tagging performance with paired bootstrap test (H_0: no difference between affix and BPE-based features)

Language	δ	p	95% CI
Welsh	0.00235	0.335	$[-0.008 \quad 0.013]$
Chichimeca	-0.01422	0.987	$[-0.027 \quad -0.001]$
Basque	0.04333	<0.001	$[0.027 \quad 0.06]$

Table 3: 1-tailed bootstrap test. H_0: using affix features and accuracy for metrics, is the multinomial logistic regression model significantly better than the perceptron model?

Language	δ	p	95% CI
Welsh	0.00164	0.358	$[-0.007 \quad 0.01]$
Chichimeca	-0.02226	1.0	$[-0.035 \quad -0.01]$
Basque	0.03484	<0.001	$[0.022 \quad 0.047]$

Table 4: 1-tailed bootstrap test. H_0: using affix features and f1 for metrics, is the multinomial logistic regression model significantly better than the perceptron model?

Language	δ	p	95% CI
Welsh	0.00846	0.116	$[-0.005 \quad 0.022]$
Chichimeca	-0.01587	0.98	$[-0.031 \quad -0.001]$
Basque	0.03281	<0.001	$[0.019 \quad 0.047]$

Table 5: 1-tailed bootstrap test. H_0: using affix features and accuracy for metrics, is the multinomial logistic regression model significantly better than the perceptron model?

Language	δ	p	95% CI
Welsh	0.00043	0.493	$[-0.013 \quad 0.015]$
Chichimeca	-0.01458	0.973	$[-0.029 \quad 0.0]$
Basque	0.02139	<0.001	$[0.011 \quad 0.031]$

Table 6: 1-tailed bootstrap test. H_0: using affix features and f1 for metrics, is the multinomial logistic regression model significantly better than the perceptron model?

4 Discussion and Conclusion

Models trained on BPE-based features either performed the same or worse than those with affix-based features. This may be because BPE, as suggested by Mager et al. (2022), does not capture morphology as effectively as linguistically-based features. It could also be due to improper tuning of BPE parameters, as we see variation in BPE performance across different languages. For example, in Welsh, BPE performed the same as the affix model. Notably, Basque is the only language where a multinomial logistic regression model outperformed a perceptron. Future research could explore why this occurs and how tuning BPE parameters like vocab size affects morphological tagging performance.

References

Batsuren, Khuyagbaatar et al. (2022). *UniMorph 4.0: Universal Morphology.* arXiv: 2205.03608 [cs.CL]. URL: https://arxiv.org/abs/2205.03608.

Gutierrez-Vasques, Ximena, Christian Bentz, and Tanja Samardi (Dec. 2023). "Languages Through the Looking Glass of BPE Compression". In: *Computational Linguistics* 49.4, pp. 943–1001. ISSN: 0891-2017. DOI: 10.1162/coli_a_00489. eprint: https://direct.mit.edu/coli/article-pdf/49/4/943/2269507/coli_a_00489.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00489.

Haley, Coleman, Edoardo M Ponti, and Sharon Goldwater (2023). "Language-Agnostic Measures Discriminate Inflection and Derivation". In: *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pp. 150–152.

Jurafsky, Daniel and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models.* 3rd. Online manuscript released January 12, 2025. URL: `https://web.stanford.edu/~jurafsky/slp3/`.

Mager, Manuel, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu (2022). *BPE vs. Morphological Segmentation: A Case Study on Machine Translation of Four Polysynthetic Languages.* arXiv: `2203.08954 [cs.CL]`. URL: `https://arxiv.org/abs/2203.08954`.

Saleva, Jonne and Constantine Lignos (Apr. 2021). "The Effectiveness of Morphology-aware Segmentation in Low-Resource Neural Machine Translation". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop.* Ed. by Ionut-Teodor Sorodoc, Madhumita Sushil, Ece Takmaz, and Eneko Agirre. Online: Association for Computational Linguistics, pp. 164–174. DOI: `10.18653/v1/2021.eacl-srw.22`. URL: `https://aclanthology.org/2021.eacl-srw.22/`.

Smit, Peter, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo (2014). "Morfessor 2.0: Toolkit for statistical morphological segmentation". In: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 21–24.

Tibshirani, Robert J and Bradley Efron (1993). "An introduction to the bootstrap". In: *Monographs on statistics and applied probability* 57.1, pp. 1–436.

A Feature Functions

This section contains pseudocode for the feature functions used. The actual code for the feature functions was written in Python, and is available on the GitHub repo.

Algorithm 1 Yield features based on affixes

```
procedure YieldAffixFeatures(word, pos)
    for $i = 1$ to 5 do
        yield $word[:i]$
        yield $word[-i:]$
    end for
    yield $pos$
end procedure
```

Algorithm 2 Yield features based on BPE

```
procedure YieldBPEFeatures(word, pos)
    $bpe\_subwords \leftarrow$ tokenize($word$)
    for all $subword \in bpe\_subwords$ do
        yield $subword$
    end for
    yield $pos$
end procedure
```