Linguistic Feature Tagging for Automatic Classification of 27 Closely-Related Quechua Varieties

Claire Benét Post and Alexis Palmer
University of Colorado, Boulder
Department of Linguistics
{benet.post,alexis.palmer} @ colorado.edu

Abstract

This paper presents a multi-dialect text classifier for Quechua that augments neural models with rule-based, linguistic information to address challenges in low-resource, morphologically complex settings. The approach is built on a carefully curated dataset spanning multiple genres, including annotated parallel bible corpora, and encodes manually annotated lexical variation and polypersonal verbal agreement as explicit features within a transformer-based classifier. Results show that neural models substantially outperform statistical baselines, enabling highly accurate multi-class classification across 27 Quechua dialects. The impact of linguistic augmentation is context-dependent: gains are minimal in high-resource settings but more pronounced in low-resource and cross-domain conditions. Overall, this work aims to contribute to the development of dialect-sensitive NLP methods for Quechua and other low-resource, morphologically rich languages.

1 Introduction

Quechua is a family of languages spoken across the Andean region, comprised of approximately 40 dialects and 9 million speakers (Adelaar, 2020; Adelaar and Muysken, 2004; Grimes, 1985; Hornberger and King, 1998). These dialects vary substantially in orthography, morphology, and syntax (Hornberger and Limerick, 2019; Limerick, 2018), creating challenges for natural language processing (NLP), particularly given the limited availability of dialect-specific resources.

Most existing NLP approaches treat Quechua as a monolithic language, collapsing dialectal distinctions into a single standardized form. As illustrated in Table 1,[1] machine translation (MT) systems like Google Translate produce hybridized morphosyntactic outputs that fail to reflect any individual variety. This dialectal homogenization not only reduces linguistic fidelity but also reinforces systemic biases in language technologies, further marginalizing underrepresented speaker communities (Blasi et al., 2022; Liu et al., 2022; Ziems et al., 2022). Similar issues arise in other indigenous language families of the Americas, such as Nahuatl and Maya, which exhibit rich dialectical variation but remain computationally underrepresented and homogenized (García et al., 2021; Riemland, 2023).

ISO	Family	"He sent me."
inb	CU	Pai-mi kacha-mu-wa-rka.
		3sg-DIR send-MOV-3sg→1sg-PST
quw	CU	Pi-mi ñuca-ra cacha-mu-ca.
		3sg-DIR 1sg-DO send-MOV-3sg→1sg.PST
qub	Q1	Pay-mi noga-ta-ga cacha-masha.
		3sg-DIR 1sg-DO-EMP send-3sg→1sg.PST
qvn	Q1	Pay-mi cachra-ra-yä-man noga-ta-ga.
		3sg-DIR send-3sg→1sg-PST-from 1sg-DO-EMP
quh	Q2	Pay-taj kacha-mu-wa-rqa.
		3sg-CON send-MOV-3sg→1sg-PST
quz	Q2	Pay-taq-mi kacha-mu-wan-pas.
		3sg-CON-DIR send-MOV-3sg→1sg-ADD
Google Translate		Pay-mi kacha-mu-wa-rqa.
		3sg-DIR send-MOV-3sg→1sg-PST

Table 1: Automatic translation of "He sent me" vs. text from manually glossed Bible corpora. Note that the MT output does not align with any particular variety, nor with any subfamily.

Even when NLP tools for Quechua exist, they are typically limited in scope, often focusing on Southern Quechua (Zevallos et al., 2022; Rios et al., 2008; Rios Gonzales and Castro Mamani, 2014), only a handful of dialects (Medina, 2013; Melgarejo et al., 2022; Vergara, 2022), or lacking dialectal specificity altogether (Chen et al., 2024; Monson et al., 2006).[2] This concentration of resources risks encoding Southern Quechua dialect norms as defaults, obscuring variation across the language

[1] Gloss explanations are in Table 8 in Appendix A.

[2] Table 9 in Appendix A provides a summary of currently available NLP tools.

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 46–63

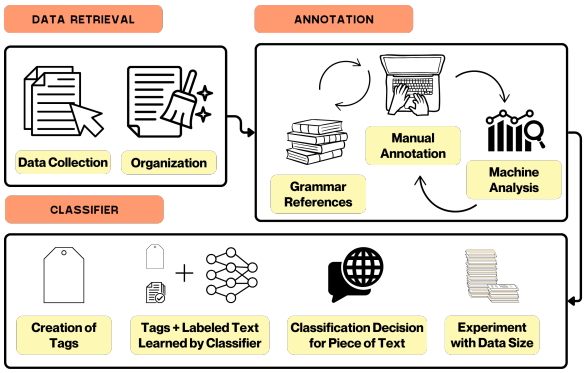

Figure 1: Dialect classification workflow.

family and limiting broader applicability. As well, the issue posed by lack of dialect-specific tools is amplified by Quechua's morphological richness, where grammatical meaning is encoded at the level of affixes. While neural models have advanced NLP significantly, purely data-driven approaches are often insufficient for capturing dialectal variation especially in morphologically rich contexts.

To address these problems, this paper adopts a rule-augmented approach that integrates neural models with linguistically informed representations on the task of dialect classification. Such approaches have shown promise in low-resource contexts by improving both performance and interpretability (Li et al., 2020; Škrlj et al., 2021; Sheth et al., 2023). Here, linguistic knowledge is incorporated not as an end in itself, but as a means of improving dialect classification. Dialect classification itself is important for being able to separate data for other downstream NLP tasks.

Thus, this paper investigates the following question: **Can linguistically informed representations improve neural dialect classification in low-resource settings?** To answer, this paper delivers the following key contributions:

1. A manually verified dataset of Quechua texts sorted by ISO dialect across multiple genres, including annotated parallel bible corpora.
2. A highly accurate multi-dialect classification framework covering 27 Quechua varieties across multiple language families.[3]
3. A systematic evaluation of linguistically informed, rule-augmented neural models, showing that their benefits are context-dependent: gains are minimal in high-resource settings but more pronounced in low-resource and cross-domain conditions.[4]
4. An exploratory analysis of segmentation strategies, showing no consistent improvements across conditions and suggesting a secondary role relative to data and feature design.

2 Related work

Early work on Quechua dialect classification is limited, with Medina (2013) providing a foundational approach using traditional machine learning methods (e.g., Naive Bayes, JRip) to distinguish between Cuzco and non-Cuzco varieties. While this work highlights challenges such as data scarcity and substantial dialectal variation, it is restricted to binary classification.

More recent advances in Quechua NLP have demonstrated the effectiveness of transformer-based models. QuBERT (Zevallos et al., 2022), a RoBERTa-based model trained on Southern Quechua, achieves strong performance on downstream tasks such as part-of-speech (POS) tagging and named entity recognition (NER). However, such models are typically trained on a limited subset of dialects and do not explicitly account for dialectal variation, limiting their applicability in multi-dialect settings.

Tokenization plays a central role in NLP for morphologically rich, low-resource languages, where effective segmentation can reduce sparsity and improve generalization. Approaches range from data-driven methods such as Byte Pair Encoding (BPE) (Shibata et al., 1999) and unigram language models (Kudo and Richardson, 2018) to linguistically informed methods like Prefix-Root-Postfix Encoding (PRPE) (Zuters et al., 2018; Chen and Fazio, 2021), which explicitly model morphological structure. Prior work shows that morphology-aware segmentation can improve downstream performance, including within the QuBERT framework (Zevallos et al., 2022).

PRPE builds on earlier work by Zuters et al. (2018) and Chen and Fazio (2021), which also explore hybrid approaches combining PRPE with BPE and unigram models for morphological neural machine translation (NMT). Additional work such as Ortega et al. (2020) proposes BPE-guided segmentation that aligns linguistic boundaries with sub-word merges, though this approach relies on

[3] All resources and code for this project may be found at the GitHub Repository: https://github.com/clairepost/Quechua_Classifier.

[4] Workflow and architecture shown in Figure 1.

ISO	Variety	No Bible	Only Bible	All Data
inb	Colombian Inga	707	151,198	151,905
qub	Huallaga Huánuco	3,350	119,909	123,259
quf	Lambayeque		160,059	160,059
quh	Southern Bolivian	214,240	135,012	349,252
quk	Chachapoyas	7,194		7,194
qul	San Martín		139,975	139,975
qup	North Bolivian		177,265	177,265
quw	Southern Pastaza		116,704	116,704
qux	Tena Lowland	203,888		203,888
quy	Yauyos	564,034	564,820	1,128,854
quz	Ayacucho	2,301,431	574,539	2,875,970
qvc	Cusco		160,816	160,816
qve	Cajamarca		167,740	167,740
qvi	Eastern Apurímac		145,704	145,704
qvh	Imbabura Highland		115,875	115,875
qvm	M-Y-L		131,047	131,047
qvn	North Junín		138,226	138,226
qvo	Napo Lowland		115,908	115,908
qvs	Huaylla Wanca		153,378	153,378
qvw	Northern Pastaza		112,921	112,921
qvz	Huaylas Ancash		157,628	157,628
qwh	Panao Huánuco	31,054	244,059	275,113
qxl	Salasaca Highland		127,034	127,034
qxh	Panao Huánuco		119,326	119,326
qxn	Northern Ancash		507,098	507,098
qxo	Southern Ancash	9,982	136,530	146,512
qxr	Cañar Highland		506,958	506,958
Total words		3,335,880	5,179,729	8,515,609
Percent		39.2%	60.8%	100%

Table 2: Word counts across datasets by ISO code and Quechua variety.

a limited set of suffixes.[5] Other preprocessing approaches include normalization pipelines for Quechua II (Rios Gonzales and Castro Mamani, 2014), though these are not always easily reproducible or generalizable across dialects.

3 Data

We collect a corpus of more than 8.5M words, across 27 varieties of Quechua as seen in Table 2.

3.1 Quechua Corpus Collection

The corpus was constructed through a combination of archival data collection, web scraping, and automated filtering. Initial data were gathered from publicly available linguistic resources, including AILLA, Runasimi, Ethnologue, Glottolog, and OLAC as well as tools such as Corpus Crawler.[6]

The data collected was cataloged with information on its source, type, and ISO code. This initial set included domains such as spoken word transcriptions (over 200k words), legal texts, and educational materials.[7] Collecting data from linguistic repositories has the advantage of providing dialectal metadata, including standard ISO code information, source location, and resource-specific dialect names.[8] These metadata were used to assign gold labels for the classifier: texts were labeled according to existing classifications in the source materials, explicit dialect names in resource titles or descriptions, or through regional information associated with the data. All dialect assignments were manually checked before inclusion in the final corpus. The final classifier evaluated in this paper was therefore trained and evaluated against manually verified dialect labels, rather than labels generated automatically by the model.

All collected data were processed to retain only Quechua content, remove non-Quechua material (e.g., Spanish and English), and ensure reliable dialect labeling. To improve data quality and obtain coherent document-level texts, a second round of collection was also conducted using a custom scraper targeting full-text Quechua bible sources across 25 dialects.

An initial version of the classifier was then used to label previously unclassified materials, enabling iterative expansion of the dataset. This process yielded a larger and more balanced corpus suitable for multi-dialect classification. The refined dataset includes texts from 27 Quechua varieties, spanning both religious and non-religious genres, with statistics shown in Table 2.

3.2 Parallel Data as Annotation Support

Additional Spanish and English bible data were collected to support cross-lingual analysis and annotation tasks. Specifically, we structured bible data into a three-way parallel format. All texts were converted into CSV files with metadata fields including `iso`, `resource`, `book`, `verse`, and `text`. Corresponding Spanish and English texts were processed in parallel, with additional linguistic annotation applied using off-the-shelf spaCy models (`es_core_news_md` and `en_core_web_md`). Alignment was performed at the verse level, allowing each Quechua segment to be paired with its Spanish and English equivalents.

To support targeted linguistic analysis, we reduce the dataset to a subset of bible texts (Matthew, Mark, and John). Selection was guided by the presence of morphosyntactic features relevant to polypersonal agreement. Specifically, a Spanish morphological parser (`es_core_news_md`) was used to identify dative and accusative clitic constructions, which often correspond to object marking in Quechua. This filtering strategy enables efficient identification of relevant constructions (Sec. 4) while maintaining cross-lingual alignment.

[5] See Table 10 in Appendix A for links to all resources.
[6] See Table 10 in Appendix A for links to all resources.
[7] See Figure 2 in Appendix A.
[8] For further information on ISO codes and family information see Table 11 in Appendix B.

While Spanish clitics provide a useful proxy for object marking, some ambiguities remain. For example, the clitic *nos* may correspond to either inclusive or exclusive first-person plural in Quechua, requiring disambiguation based on context and reference to dialect-specific grammars. Third-person clitics (e.g., *le*, *lo*, *la*) were not explicitly targeted due to their limited role in object agreement, though some instances were retained when co-occurring with other relevant features.

For each classification experiment, data were split into training and evaluation sets using an 85/15 partition at the level of text chunks. Documents were segmented into overlapping chunks of 250 words (with 50-word overlap), and these chunks served as the unit of classification. Splits were performed randomly without stratification, resulting in class distributions that reflect the natural imbalance of the dataset. The full-data setting includes 27 dialects, while the no-bible condition contains only 10 dialects.

4 Annotation

This section outlines the linguistic annotation and feature development used to support dialect classification. We consider: (i) lexical variation across dialects, (ii) polypersonal verbal agreement, and (iii) construction of a morphological inventory used to derive segmentation heuristics.[9] Together, these annotations inform both feature design and preprocessing for the classifier.

Some forms are used across multiple dialects. When counting, we consider both the number of **unique** surface forms, regardless of dialect, and the **total** number of terms identified for all dialects, which includes duplicated surface forms. The total number of collected items for each of these annotation sets is:

- **Lexical terms:** 86 unique, 330 total
- **Polypersonal morphemes:** 110 unique, 157 total
- **Additional morphemes:** 448 total

4.1 Lexical Variation

Our approach to analyzing lexical variation across Quechua dialects follows Medina (2013), who identify differences in core vocabulary items across varieties. We extend her analysis to additional dialects using both corpus data and dialect-specific grammatical resources. Lexical items under comparison include translations for terms (referred to here by their English words) such as *father*, *land*, *sun*, and *moon*. We exclude items exhibiting high uniformity across dialects (e.g., *mother* = *mama*, *water* = *yacu*, *woman* = *warmi*), due to their limited diagnostic value.

For each dialect, we identify the relevant lexical forms through corpus search, cross-referenced with dictionaries and grammars. When multiple lexical variants are attested for a single concept-variety pair, we record all variants (e.g., *tayta/taita* for *father*). These variants are compiled into a comparative lexical chart (Figure 3 in Appendix B), enabling systematic comparison across varieties.

We supplement the collected lexical data through extensive consultation of dialect-specific resources: a grammar of Huallaga Huánuco Quechua (Weber, 1989), a dictionary of Margos-Yarowilca-Lauricocha Quechua (Bean, 1986), a grammar of North Junín Quechua (Adelaar, 1977), a grammar of Huaylla Wanca Quechua (Cerrón-Palomino, 1976), a dictionary of Huaylas Ancash Quechua (Carranza Romero, 2003), a dictionary of Panao Huánuco Quechua (Smith, 1994), a grammatical sketch of Northern Conchucos Ancash Quechua (Wroughton, 1988), a grammar sketch of Southern Ancash Quechua (Hintz, 2017), and a dictionary of Ayacucho Quechua (Parker, 1969). In addition to core vocabulary, pronominal roots encoding person and number distinctions (`1SG`, `2SG`, `3SG`, `1PL.INCL`, `1PL.EXCL`, `2PL`, `3PL`) were also incorporated to ensure broader lexical coverage across dialects.

4.2 Polypersonal Verbal Suffixes

One of the most salient morphosyntactic features across Quechua dialects is *polypersonal verb agreement*, in which a single verb encodes both subject and object through suffixation. This system, widely documented in typological and descriptive work (Adelaar and Muysken, 2004; Lakämper and Wunderlich, 1998; Camacho Rios, 2020; Rataj, 2015), reflects the agglutinative structure of Quechua and provides a rich source of dialectal variation.

To capture this variation, we manually annotate polypersonal constructions across 25 Quechua varieties using the collected parallel bible corpora.[10] For each sentence-level instance, annotations include subject and object person/number (e.g., `1SG`

[9] More information on ISO codes and dialects in Table 11 in Appendix B.

[10] See Figure 4 in Appendix B.

Family	Added Morphs
Quechua I	+144
Colombia-Ecuador	+92
Cajamarca-Lambayeque	+70
San Martín-Amazonas	+62
Quechua II	+80

Table 3: Morphemes added to PRPE per Quechua family

`Subj → 2PL Obj`), verb stem, and relevant suffixes, along with aligned Spanish and English translations. Given the complexity of Quechua verbal morphology, annotation focuses primarily on present indicative forms, which most consistently encode subject–object agreement. Ambiguous cases (particularly those involving tense, mood, or aspect distinctions) are excluded from final analyses when reliable interpretation is not possible. Reference grammars and linguistic descriptions were consulted extensively to support annotation decisions,[11] particularly for identifying object-marking suffixes and "transitions" between subject and object forms. When necessary, additional resources such as SAILS[12] were used to confirm morphological patterns. As an example, we use this resource to determine that in qvn, the affix for 1sgQB is $-wa$.

Cross-dialect comparison reveals some typological trends. Conservative dialects, particularly those spoken in highland regions of Peru and Bolivia, tend to preserve rich agreement paradigms, including distinct markers for first- and second-person objects. In contrast, dialects in Ecuador and Colombia often display morphosyntactic simplification, with reduced or absent object agreement marking. These patterns provide informative features for dialect classification.

4.3 Morphological Inventory and Heuristics

To support linguistically informed segmentation, we construct an expanded inventory of Quechua morphemes through manual analysis of grammatical resources. While some grammars provide explicit morpheme lists, many require close examination of descriptive text to identify suffixes and inflectional patterns, as inconsistent formatting and multi-language scripts often render OCR ineffective. This process involved extracting both common and dialect-specific morphemes, including case markers, tense and aspect suffixes, evidentials, and derivational affixes. The resulting inventory extends prior PRPE resources (Chen and Fazio, 2021) and forms the basis for dialect-aware segmentation heuristics used in subsection 5.1.

We derive heuristics from this inventory by treating morphs marked with a hyphen (e.g., `-nchi`) as suffixes and those without as roots. Given that Quechua lacks productive prefixation, suffix-based segmentation is particularly central. Morphemes are first grouped by dialect and then consolidated into family-level inventories corresponding to major Quechua families: Colombia-Ecuador Quechua (cu: `qxl, inb, qup, quw, qvi, qvo, qvz, qxr`), Cajamarca-Lambayeque Quechua (ca: `qvc, quf`), Quechua I (q1: `qub, qux, qvh, qvm, qvn, qvw, qwh, qxh, qxn, qxo`), Quechua II (q2: `quy, quh, qul, quz, qve`), and San Martín-Amazonas Quechua (sm: `qvs, quk`).

Morphemes with multiple meanings (e.g., `-pi`) are counted only once per family. Compared to the 64 morphemes used in Chen and Fazio (2021), this work substantially expands coverage across all families, adding a total of 448 morphemes (Table 3).

5 Methodology

We investigate a range of modeling strategies, varying both data settings and, crucially, degree to which linguistic information is incorporated into the model. This section describes modeling options, and we present results in Section 6.

5.1 Segmentation Strategies

To evaluate the impact of input representation on dialect classification, we explore multiple segmentation strategies, including BPE, unigram language models (Kudo and Richardson, 2018), PRPE (Zuters et al., 2018; Chen and Fazio, 2021), and a hybrid PRPE+BPE approach (Chen and Fazio, 2021; Ortega et al., 2020).

We extend PRPE with the dialect-informed morphological inventory described in subsection 4.3, enabling segmentation to better capture variation across Quechua families. These heuristics guide suffix identification and improve alignment between sub-word units and linguistically meaningful morphemes.

In addition, we train our segmentation at the level of major Quechua language families, allowing representations to capture shared morphological patterns within families while preserving cross-dialect distinctions. This setup enables controlled comparison of how segmentation and linguistic granularity affect classification performance.

[11] See Table 12 in Appendix B.

[12] https://sails.clld.org/languages/qvc

5.2 Model Architectures

QuBERTa Classifier. Our primary neural classifier is built on QuBERTa, a RoBERTa-based model adapted for Quechua (Zevallos et al., 2022). Input texts are preprocessed by removing numeric tokens and segmenting documents into overlapping chunks (250 tokens with 50-token overlap) to satisfy the model's 512-token input constraint. Chunks are tokenized and padded using the `Llamacha/QuBERTa` tokenizer.

The model is fine-tuned for dialect classification using standard optimization procedures, with hyperparameters (batch size, learning rate, number of epochs) tuned on validation data. We evaluate performance with standard classification metrics.

Rule-Augmented QuBERTa. Our rule-augmented variant extends the base QuBERTa model by incorporating linguistically motivated features derived from lexical and polypersonal annotation. We encode these features as tags and append them to each text chunk during preprocessing, allowing the model to access explicit morphosyntactic and lexical cues alongside learned representations.

Lexical tags are generated by matching dialect-specific vocabulary items and appending corresponding markers (e.g., `<TAG_tayta>`). Polypersonal tags are derived through suffix matching over a curated list of verbal affixes, using a longest-match strategy to prioritize more specific morphological forms. Identified polypersonal verbal suffixes are encoded as tags (e.g., `<VERB_wanku>`) and added as additional tokens to the input.

We use the same training procedure for both the base model (standard QuBERTa classifier) and the rule-enhanced version, allowing direct comparison of performance with and without linguistic feature injection. Separate configurations evaluate the contribution of lexical tags, polypersonal tags, and their combination.

Naive Bayes Baseline. We implement a Multinomial Naive Bayes classifier as a statistical baseline, following Medina (2013). The model operates over TF-IDF representations, which encode term frequency and corpus-level importance to capture distributional patterns in word usage.

To ensure comparability with the neural models, the same cleaned and chunked inputs are used. TF-IDF features are extracted from each text segment and used to train a multi-class classifier over dialect labels. To mitigate class imbalance across dialects, a limit is placed on the number of chunks per document, preventing over-representation of high-resource sources. This constraint after experimentation proved to improve stability and provides a more balanced comparison with neural approaches.

Model	Rules *(# Dialects)*	All *(27)*	NO BIBLE *(10)*	ONLY BIBLE *(25)*
Bayes	No rules	**.9385**	**.7551**	**.9713**
Bayes	Lexical+Verb	.9064	.7506	.9406
Neural	No rules	.9907	.9662	.9957
Neural	Lexical	.9916	.9664	.9949
Neural	Verb	**.9936**	**.9670**	**.9969**
Neural	Lexical+Verb	.9928	.9643	.9961

Table 4: Weighted F1 scores across models, rules, and data settings with standard QuBERTa tokenization.

5.3 Experimental Settings

To evaluate model performance under varying resource conditions, we compare several data settings. In the high-resource ALL setting, all available corpora (see Table 2) from 27 dialects are included, comprising religious texts, other written materials, and previously unclassified data labeled through the pipeline described in Section 5.2. This setting is relatively balanced across varieties and domains. In the limited-data setting (NO BIBLE, or NB), bible data are removed to simulate a low-resource and domain-mismatched environment. This results in a reduced set of 10 dialects with highly imbalanced distributions, where resource availability ranges from large corpora (over 2 million words) to very limited data (fewer than 10,000 words). A third configuration utilizing only bible data (ONLY BIBLE, or B) is confined to a single domain, with data for 25 dialects. This corpus is close to fully-balanced; for some dialects, we have the complete bible and for others only the New Testament.

Model performance is measured using standard classification metrics, including accuracy, precision, recall, and F1-score, along with confusion matrices to analyze dialect-specific errors.[13]

6 Results

6.1 Overall Performance

Overall results across all experimental conditions are summarized in Table 4. The QuBERTa-based models consistently outperform the statistical baseline (modeled after Medina (2013)) across all set-

[13] See Appendix C.

tings, maintaining F1 scores above .96 even in the most constrained condition, compared to approximately .75 for Naive Bayes.

In the ALL setting, the base QuBERTa classifier achieves near-ceiling performance without linguistic augmentation (.9907 F1). Rule-based features yield only marginal gains, with the best configuration (verb rules) reaching .9936 F1.[14] Statistical comparison confirms that this improvement is small and not significant (ΔF1 = +0.0015, $p = 0.133$).[15] Similar patterns hold in the ONLY BIBLE setting, where performance remains uniformly high due to domain homogeneity.

In the limited-data setting (NO BIBLE), the neural baseline drops to .9662 F1. Rule augmentation produces more variable effects, with the best-performing model (verb rules) reaching .9670 F1. While average per-dialect improvements are larger (ΔF1 = +0.0649), high variance results in non-significant tests ($p = 0.398$). Notably, the weighted change is slightly negative, suggesting gains are concentrated in lower-resource dialects (e.g., qvo, qwh) rather than uniformly distributed.

quz	quh	English
<hallp'a>	<pacha>	earth
<inti>	<indi>	sun
<quilla>	<killa>	moon
<runa>	<ullku>	person

Table 5: Dialectal variation encoded through lexical document-level tagging.

Examination of misclassification patterns shows a concentration of errors among closely related Southern Quechua varieties.[16] The baseline no-rules model frequently confuses Cuzco Quechua (quz) with Southern Bolivian Quechua (quh), while the rules-augmented model reduces these errors, possibly by leveraging fine-grained lexical distinctions between texts, such as those seen in Table 5.[17]

Averaging across languages, classification performance is consistently high. This setting, though, assumes availability of significant amounts of data across varieties. To get a more nuanced picture of performance, we next investigate performance in more realistic settings.

[14] Further details on morpheme exclusion in the verb rules are provided in Appendix C.

[15] See Table 14 in Appendix C.

[16] See Table 11 in Appendix A.

[17] For a confusion matrix comparison see Table 15 and Table 16 in Appendix C.

Data settings				*Rule settings*			
Train	**#**	**Eval**	**#**	**No**	**Lex**	**Verb**	**Lex+V**
NB	10	ALL	27	.4032	.3805	**.4039**	.3936
NB	10	ALL-F	27	.4999	.4429	**.5344**	.4754
NB	10	ALL ∩ NB	10	.7646	.7663	**.7745**	.7701
NB	10	ALL ∩ NB-F	10	.8328	.8360	**.8424**	.8303
B	25	NB	10	.7105	**.8094**	.8054	.8038
B	25	NB ∩ B	8	**.8288**	.8283	.8195	.8227
B ∩ NB	8	NB	10	**.7594**	.7587	.7556	.7526
B ∩ NB	8	NB ∩ B	8	**.8182**	.8015	.8171	.8127

Table 6: Cross-domain weighted F1 results across training and evaluation conditions. # is number of dialects per setting. (-F) indicates family level evaluation. NB/B= NO BIBLE/BIBLE ONLY data. $X \cap Y$ represents the intersection of X and Y datasets.

6.2 Cross domain experiments

To assess generalization, we conduct cross-domain experiments with training and evaluation data from different distributions (e.g., NB vs. B). These settings better reflect real-world conditions where domain mismatch is common, especially in low resource settings for Indigenous languages. Cross-domain performance (Tab. 6) is substantially lower than in-domain results (Tab. 4), reflecting differences in lexical choice, genre, and dialect coverage.

We first examine a low-resource model trained on NB (10 dialects) and evaluated on ALL (27 dialects). Because of the mismatch between labels, we introduce a relaxed evaluation condition (-F), counting family-level matches as correct (groupings in App. B.1). This partially compensates for missing dialect coverage, though some families remain absent in training data (e.g., San Marín).

Under this setting, performance largely degrades, but differences emerge across rule configurations. In the relaxed condition, the no-rules model achieves .4999 F1, while the verb-augmented model improves to .5344, representing the largest gain observed in this configuration. This suggests that verb-based features provide useful structure when generalizing beyond the training distribution. Across dialects, the mean improvement is positive (ΔF1 = .0511, $d = 0.21$), though not statistically significant (Wilcoxon $p = 0.50$).[18]

Restricting evaluation to dialects present in the training set (10 dialects) improves performance across all configurations (F1 $> .76$), indicating that much of the degradation is driven by label space mismatch. In this controlled setting, verb rules still provide a small but consistent improvement, and provide more evidence that linguistic features are

[18] See Table 14 in Appendix C.

particularly valuable when models must generalize beyond their training label space, rather than simply interpolate within it when moving from a smaller model setting.

The final set of experiments trains on the larger but domain-restricted BIBLE ONLY corpus and evaluates on heterogeneous NO BIBLE data. Performance again drops substantially, indicating that bible text does not provide a fully representative training distribution. This observation aligns with prior work: although bible corpora are widely used due to accessibility and multilingual coverage (Christodouloupoulos and Steedman, 2015), they reflect formal or archaic registers, translation-driven inconsistencies, and a narrow range of genres (Levshina, 2022; Hutchinson, 2024).

Despite this domain mismatch, rule augmentation improves performance for the no-rules model, which achieves .7105 F1, while lexical tagging performs best (.8094), outperforming verb (.8054) and combined rules (.8038). This finding stands in contrast to prior experiments, in which verb-based rules yielded the highest effectiveness, thereby suggesting that the utility of specific linguistic features depends on the direction of the domain shift.

One explanation is that BIBLE ONLY training already exposes the model to relatively consistent verbal morphology, reducing the added value of verb-based rules at test time. Lexical tagging instead helps compensate for lexical and genre differences between bible and non-bible data. This is supported by per-dialect results, where gains are concentrated in higher-support dialects such as *quz* (F1 .7525 no rules to .8894 verb rules), while many low-resource dialects show minimal change.[19]

Looking at the bottom of Table 4, when evaluation is restricted to overlapping dialect subsets, though, differences between rule configurations are negligible (e.g., .7594 vs. .7587). Thus the effectiveness of linguistic augmentation seems to depend not only on domain shift, but also on label space alignment and data coverage.

6.3 Effect of Segmentation

Segmentation strategy does not produce consistent improvements across conditions. While morphology-aware approaches (e.g., PRPE+BPE) show gains in some low-resource settings, these effects are not stable (see Table 7). In high-resource settings, differences between segmentation methods are minimal and models seem to have sufficient data to learn effective representations regardless of segmentation strategy. In lower-resource conditions, results are more variable, with no single method consistently outperforming others.[20]

Model	Segment	Low Data		All Data	
		F1w	F1m	F1w	F1m
ca	bpe	**.945**	.533	**.972**	**.935**
ca	prpe	.933	.468	.964	.922
ca	prpe+bpe	.931	.454	.971	.930
ca	unigram	.941	**.534**	.967	.929
q1	bpe	.937	.485	**.978**	.942
q1	prpe	.937	.458	.977	.941
q1	prpe+bpe	**.949**	.534	.977	**.943**
q1	unigram	.948	**.560**	**.978**	.941
q2	bpe	.966	.644	.986	**.967**
q2	prpe	**.968**	.649	**.988**	.952
q2	prpe+bpe	**.968**	**.663**	.987	.951
q2	unigram	.967	.626	**.988**	.963
sm	bpe	.923	.384	**.960**	**.920**
sm	prpe	.923	.429	.945	.903
sm	prpe+bpe	**.929**	**.438**	.939	.895
sm	unigram	.921	.395	.955	.913

Table 7: Segmentation results grouped by dialect family under low-resource and full-data settings. F1w = weighted F1 (accounts for class imbalance), F1m = macro F1 (equal weighting across dialects).

7 Conclusion & Future Work

This paper presents a multi-dialect classification framework for Quechua that augments neural models with linguistically informed features. Across all settings, neural models substantially outperform statistical baselines, enabling accurate multi-class classification across 27 Quechua dialects.

The impact of linguistic augmentation, however, is nuanced. In high-resource and homogeneous (bible) settings, performance is already near ceiling, and rule-based features provide only marginal gains. In contrast, in low-resource and cross-domain conditions, linguistic features become more valuable, though their effects are uneven and depend on both data availability and evaluation setup. In particular, verb-based features are most beneficial when generalizing from limited training data to broader label spaces, while lexical tagging was found to be more effective under domain shift, especially when transferring from bible-only to heterogeneous corpora. The utility of linguistic features then may be context-dependent rather than uniformly additive.

More broadly, this work aims to contribute to on-

[19] For a confusion matrix comparison see Table 18 and Table 17 in Appendix C.

[20] A segmentation results breakdown is in Appendix C.

going efforts to integrate linguistic knowledge into modern machine learning pipelines. By advancing dialect-sensitive NLP tools, it hopes to support more accurate and inclusive language technologies, helping to address the marginalization of Quechua speakers in digital spaces.

Future work will extend this approach in several directions. First, we plan to develop family-specific RoBERTa models that better capture variation beyond Southern Quechua. Second, we aim to expand and refine morphological inventories used in segmentation and tagging. Additional directions include evaluating segmentation in downstream morphological tasks and revisiting alternative sub-word strategies such as BPE-guided methods (Ortega et al., 2020). Third, this approach would greatly benefit from expansion beyond bible corpora, and work is underway to annotate more varied domains across additional dialects. Finally, for applications to other language families, our results suggest that lexical tagging offers the most favorable trade-off between annotation effort and performance gains, particularly in cross-domain and low-resource settings.

Limitations

A limitation of this work is the reliance on domain-specific data, particularly bible texts, which constitute a large portion of the available corpora. While we explicitly evaluate a NO BIBLE setting to simulate low-resource conditions, the properties of religious text may not fully reflect naturalistic language use and might have some effect on classification of non-bible texts. Current work is aimed at expanding the corpus to more diverse domains for a wider range of dialects.

As well, although this work expands classification to 27 varieties of Quechua, there are still more within the language family. Varieties were included if *any* data could be found and preprocessed into text files. Even still, this leaves room for improvement by finding additional textual resources.

Ethics

This work involves the use of textual data from various publicly available sources for different Quechua dialects. We recognize the importance of indigenous data sovereignty and the need to handle language data in a way that respects the communities from which it originates. Wherever possible, this work relies on publicly available sources and keeps record of metadata and authorship in order to further support language documentation efforts.

At the same time, automated dialect classification systems carry potential risks as misclassifications may obscure dialectical distinctions, reinforce inaccurate generalizations, or privilege better-represented varieties over lower-resource ones. These risks are particularly relevant in cross-domain settings, where model performance is less stable and uneven across dialects. As such, the models presented here should not be used as authoritative tools for linguistic identification, but rather as assistive technologies whose outputs require careful interpretation and are up for welcome debate by community members.

Acknowledgments

Most sincere thanks to the two anonymous reviewers for thorough, insightful, and helpful reviews. This work would not have been possible without the help and advice of our colleagues. First, thanks to Wilma Doris Loayza for her Quechua teaching materials that inspired the study into polypersonal verbal agreement patterns for this project. Next, the modeling portion of this paper was assisted by Saksham Khatwani, who helped give feedback on the classifier and advice on implementing the different segmentation methods. Additional thanks are in order to Andrew Cowell, Hannah Haynie, and Jim Martin for their feedback at various points throughout this project. Parts of this work were supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another."

References

Willem Adelaar. 1977. *Tarma Quechua: Grammar, texts and dictionary*. Ph.D. thesis, Universiteit van Amsterdam, Lisse.

Willem F. H. Adelaar and Pieter C. Muysken. 2004. *The Inca Sphere*, page 165–410. Cambridge Language Surveys. Cambridge University Press.

Willem FH Adelaar. 2020. Morphology in Quechuan Languages. In *Oxford Research Encyclopedia of Linguistics*.

Mark Bean. 1986. *Pequeño diccionario de palabras útiles: Quechua-castellano, castellano-quechua*, primera edición edition. Dirección Departmental de Educación - Huánuco e Instituto Lingüístico de Verano, Perú. Published. Available online: https://www.sil.org/resources/archives/27945.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Gladys Camacho Rios. 2020. *Verb morphology in South Bolivian Quechua: A case study of the Uma Piwra rural variety*. Ph.D. thesis.

Lawrence Kidd Carpenter. 1982. *Ecuadorian Quichua: Descriptive sketch and variation*. University of Florida.

Francisco Carranza Romero. 2003. *Diccionario quechua ancashino-castellano*. Iberoamericana.

Rodolfo Cerrón-Palomino. 1976. *Gramática quechua: Junín/Huanca*. Ministerio de Educación/IEP, Lima.

Junhao Chen, Peng Shu, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Zhengliang Liu, Lewis C Howe, and Tianming Liu. 2024. QueEn: A Large Language Model for Quechua-English Translation. *arXiv preprint arXiv:2412.05184*.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Luis Cordero Crespo. 1955. *Diccionario de Quichua*. Biblioteca Hernán Malo González.

S Khalil Bello García, E Sánchez Lucero, E Bonilla Huerta, J Crispín Hernández Hernández, J Federico Ramírez Cruz, and B Estela Pedroza Méndez. 2021. Implementation of neural machine translation for Nahuatl as a web platform: a focus on text translation. *Programming and Computer Software*, 47:778–792.

Joseph E Grimes. 1985. The interpretation of relationships among quechua dialects. *Oceanic Linguistics Special Publication*, pages 271–284.

Manuel Guzmán. 1920. *Gramatica de la lengua Quichua (dialecto del Ecuador)*. Prensa Católica, Quito.

Marleen Haboud de Ortega, Luis Montaluisa Chasiquiza, Fabián Muenala Pineda, and Froilan Viteri Gualinga. 1982. *Shimiyuc-Panca, Caimi Ñucanchic*. Pontificia Universidad Católica and Ministerio de Educación y Cultura, Quito. Yanapaccuna: Filemón Aguinda Díaz, Mariano Cerda Chimbo, Enrique Contreras Ponce, Manuel Díaz Cajas, Agustín Jérez Jérez, Luis Macas Ambuludi, César Shiguango Grefa, Consuelo Yánez Cossío. Shuyuccuna: José Aviléz López, José Higuera Rosero. Quillcac: Humberto Cachihuango, Manuel Serrano, Gladys Muenala Vega.

Daniel J. Hintz. 2017. *El Aspecto Verbal en Quechua Campos Semánticos Entretejidos y el Surgimiento de Sistemas Gramaticales*, volume 58 of *Serie Lingüística Peruana*. Instituto Lingüístico de Verano, Lima.

Nancy H Hornberger and Kendall A King. 1998. Authenticity and unification in Quechua language planning. *Language Culture and Curriculum*, 11(3):390–410.

Nancy H Hornberger and Nicholas Limerick. 2019. Teachers, textbooks, and orthographic choices in Quechua: Bilingual intercultural education in Peru and Ecuador. In *Perspectives on Indigenous writing and literacies*, pages 141–164. Brill.

Ben Hutchinson. 2024. Modeling the sacred: Considerations when using religious texts in natural language processing. *arXiv preprint arXiv:2404.14740*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Renate Lakämper and Dieter Wunderlich. 1998. Person marking in Quechua—A constraint-based minimalist analysis. *Lingua*, 105(3-4):113–148.

Natalia Levshina. 2022. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.

Tao Li, Parth Anand Jawale, Martha Palmer, and Vivek Srikumar. 2020. Structured tuning for semantic role labeling. *arXiv preprint arXiv:2005.00496*.

Nicholas Limerick. 2018. Kichwa or Quichua? Competing alphabets, political histories, and complicated reading in Indigenous languages. *Comparative Education Review*, 62(1):103–124.

Zoey Liu, Crystal Richardson, Richard Hatcher Jr, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. *arXiv preprint arXiv:2204.05541*.

Rosemary Jiménez Medina. 2013. *Clasificación Por Dialecto De Documentos Escritos En Quechua*. Ph.D. thesis, Universidad Nacional De San Antonio Abad.

Nelsi Melgarejo, Rodolfo Zevallos, Héctor Gómez, and John E Ortega. 2022. WordNet-QU: Development of a lexical database for Quechua varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433.

Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. 2006. Building NLP systems for two resource-scarce indigenous languages:

Mapudungun and Quechua. *Strategies for developing machine translation for minority languages*, page 15.

Carolyn Orr. 1973. *Dialectos quichuas del Ecuador con respecto a lectores principiantes*. ILV, Quito.

Carolyn Orr. 1992. *Runa shimi (Gramática quichua de Tena)*. ILV, Quito.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Gary John Parker. 1969. *Ayacucho Quechua grammar and dictionary*, volume 82 of *Janua linguarum : Series practica*. Mouton, The Hague. Bibliography: p. [226].

Félix Quesada. 1976. *Gramática quechua: Cajamarca-Cañaris*. Ministerio de Educación. Instituto de Estudios Peruanos.

J. F. M. Raez. 1917. *Gramáticas en el quichua-huanca y en el de Ayacucho*. Sanmarti y Ca, Lima.

Vlastimil Rataj. 2015. Marcación de la segunda persona objeto en las transiciones del quechua cusqueño. *IBERO-AMERICANA PRAGENSIA*, 43(1):27–46.

Matt Riemland. 2023. Theorizing sustainable, low-resource MT in development settings: Pivot-based MT between Guatemala's indigenous Mayan languages. *Translation Spaces*, 12(2):231–254.

Annette Rios. 2010. Applying finite-state techniques to a native american language: Quechua. *Institut für Computerlinguistik, Universität Zürich*.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A Quechua-Spanish parallel treebank. *LOT occasional series, Netherlands Graduate School of Linguistics*, 12:53–64.

Annette Rios Gonzales and Richard Alexander Castro Mamani. 2014. Morphological disambiguation and text normalization for Southern Quechua varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 39–47, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Amit Sheth, Kaushik Roy, and Manas Gaur. 2023. Neurosymbolic artificial intelligence (why, what, and how). *IEEE Intelligent Systems*, 38(3):56–62.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Blaž Škrlj, Matej Martinc, Nada Lavrač, and Senja Pollak. 2021. autoBOT: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, 110(5):989–1028.

Terrence Smith. 1994. *Alli rimay ashina (Un pequeño diccionario de palabras utiles en el quechua de Panao)*. Dirección Regional de Educación-Huánuco and Instituto Lingüístico de Verano, Huánuco.

Nelsi Belly Melgarejo Vergara. 2022. Desarrollo de recursos léxicos multi-dialécticos para el quechua. Master's thesis, Pontificia Universidad Catolica del Peru (Peru).

David Weber. 1989. *A grammar of Huallaga (Huánuco) Quechua*, volume 112. Univ of California Press.

James F. Wroughton. 1988. Major Clause Constituents of Conchucos (Ancash) Quechua. Master's thesis, University of Texas at Arlington, Arlington.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Yoshikawa, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. *arXiv preprint arXiv:2204.03031*.

Jānis Zuters, Gus Strazds, and Kārlis Immers. 2018. Semi-automatic quasi-morphological word segmentation for neural machine translation. In *International Baltic conference on databases and information systems*, pages 289–301. Springer.

A Appendix: Background

Gloss	Meaning
1sg	first person singular
3sg	third person singular
DIR	direct evidential
DO	direct object marker
MOV	directional motion away
EMP	emphatic marker
CON	contrastive marker
PST	past tense
ADD	additive marker ('also')
3sg→1sg	third person subject acting on first person object

Table 8: Gloss abbreviations explanations for the morphological parses in Table 1.

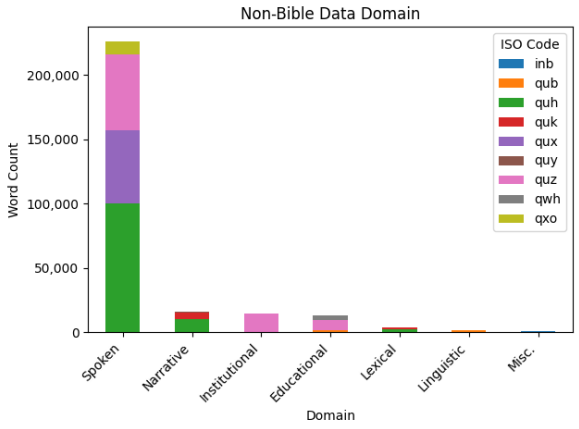

Figure 2: Word counts broken up by domain type and ISO code for non-bible data corpora.

B Appendix: Data

B.1 Dialect grouping

The dialect groupings used in this study follow the family hierarchy and clade structure outlined by Glottolog.[21] This classification allows for hierarchical grouping of varieties at the family, subfamily, and dialect levels. For instance, dialects from the Quechua I group (e.g., Cajamarca Quechua) are modeled separately from those in Quechua II-A or II-B (e.g., Ayacucho or Ancash Quechua). Dialect identity and family membership were assigned using this system because they are widely used by other resources as classification for data. A table of family composition, summarizing clade-level distinctions and vocabulary comparisons, is shown in Table 11.

[21] https://glottolog.org/resource/languoid/id/quec1387

Dialects	NLP Tools
Cuzco Quechua	Text classifier (Medina, 2013), AntiMorfo: finite state transducer morphological analyzer (Rios, 2010)
Southern Quechua	QuBERT: monolingual BERT model (Zevallos et al., 2022), SQUOIA: finite state transducer morphological analyzer & spell checker (Rios Gonzales and Castro Mamani, 2014), Quechua-Spanish Treebank (Rios et al., 2008)
Southern, Central, Northern, and Amazonian Quechua	WordNET-QU: lexical database (Melgarejo et al., 2022), POS tagging models (Vergara, 2022)
Unspecified	QueEn: LLM machine translation (Chen et al., 2024), Quechua-Spanish machine translation & morphological analyzer (Monson et al., 2006)

Table 9: Quechua Dialects and Corresponding NLP Tools

B.2 Linguistic Annotation

Drawing from the Medina (2013) analysis of lexical differences across Quechua dialects, a thorough examination of vocabulary discrepancies was conducted, as shown in Figure 3. This included dialects not originally covered by Medina, utilizing both corpus data and dialect-specific grammar resources. If there were multiple words that were used for a lexical term, we indicated that with a / in the chart.

As well, Figure 4 gives a summary of all the polypersonal verbal suffixes found during the verbal annotation phase. Grammars used for help in finding specific polypersonal verbal agreement information is seen in Table 12.

Resource	Link
Ortega et al. (2020)	https://github.com/johneortega/mt_quechua_spanish/tree/master/bpe_guided
AILLA	https://ailla.utexas.org
Runasimi	https://runasimi.de
Ethnologue	https://www.ethnologue.com
Glottolog	https://glottolog.org
OLAC	http://www.language-archives.org
Bible data	https://live.bible.is
QuBERTa tokenizer	https://huggingface.co/Llamacha/QuBERTa

Table 10: Resources and links

				Lexical Terms					
ISO	Dialect Name	Family	Variety	father	land	water	sun	moon	man
qub	Huallaga Huánuco Quechua	Quechua 1	Central Quechua 1	tayta	allpu	yacu	inti	quilla	runa
qux	Yauyos Quechua	Quechua 1	Yauyosic	tayta	pacha	yaku	rupay/inti	killa	runa
qvh	Huamalíes-Dos de Mayo Huánaco Quechua	Quechua 1	Central Quechua 1	tayta	pacha	yaku	inti	killa	runa
qvm	Margos-Yarowilca-Lauricocha Quechua	Quechua 1	Central Quechua 1	tayta	patsa	yacu	inti	quilla	runa
qvn	North Junín Quechua	Quechua 1	Central Quechua 1	tayta	pacha/alpa	yacu	inti	quilla	runa
qvw	Huaylla Wanca Quechua	Quechua 1	Central Quechua 1	taytä	pacha/allpa	yacu	inti	quilla	nuna
qwh	Huaylas Ancash Quechua	Quechua 1	Central Quechua 1	yaya/papä	pacha/allpa	yacu	inti	quilla	nuna
qxh	Panao Huánuco Quechua	Quechua 1	Central Quechua 1	tayta	pacha	yacu	inti	quilla	runa
qxn	Northern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	tayta	patsa	yacu	rupay	quilla	runa/ollgo
qxo	Southern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	tayta	patsa	yacu	rupay	quilla	runa
quy	Ayacucho Quechua	Quechua 2	Ayacuchan Quechua	tayta	pacha	yaku	inti	killa	runa
quh	Southern Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua	tata	pacha	yaku	inti	kuilla	runa
qul	North Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua	tata	pacha	unu	inti	killa	runa
quz	Cusco Quechua	Quechua 2	Cuscan Quechua	tayta	hallp'a	unu	inti	killa	qhari
qve	Eastern Apurímac Quechua	Quechua 2	Cuscan Quechua	tayta/papa	pacha/allpa	unu	inti	killa	qari
qvc	Cajamarca Quechua	Cajamarca-Lambayeque Quechua	Cajamarca Quechua	tayta	allpa/pacha	yaku	rupay	killa	runa
quf	Lambayeque Quechua	Cajamarca-Lambayeque Quechua	Lambayeque Quechua	tayta	pacha	yaku	rupay	killa	runa
qxl	Salasaca Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua A	yaya/tayta	pacha	yaku	indi	killa	kari
inb	Colombian Inga	Colombia-Ecuador Quechua	Ecuadorian Quechua B	taita	patsa	iaku	indi	killa	runa/kari/jinti
qup	Southern Pastaza Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	allpa	yaku	inti	killa	kari
quw	Tena Lowland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	allpa	yacu	inti	quilla	runa
qvi	Imbabura Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	jahua/papá	pacha	yacu	inti	quilla	runa
qvo	Napo Lowland Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	pacha/allpa	yacu	inti	quilla	runa
qvz	Northern Pastaza Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	allpa/pacha	yacu	indi	quilla	runa
qxr	Cañar Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	taita	pacha	yacu	inti	quilla	runa
qvs	San Martín Quechua	San Martin-Amazonas Quechua	Chachapoyas Quechua	tata	pacha	yaku	inti	killa	runa
quk	Chachapoyas Quechua	San Martin-Amazonas Quechua	Chachapoyas Quechua	tata	allpa	yaku	inti	killa	runa/ullku

Figure 3: Common words for each dialect organized by Quechua families and varietal sub-families.

						Verbs															
						1sg object				1pl.incl object	1pl.excl object				2sg object				2pl object		
ISO Code	Dialect Name	Family	Clade 1	Clade 2	Clade 3	2sg1sg	3sg1sg	2pl1sg	3pl1sg	3sg1pl.incl	2sg1pl.excl	3sg1pl.excl	2pl1pl.excl	3pl1pl.excl	1sg2sg	3sg2sg	1pl.excl2sg	3pl2sg	1sg2pl	3sg2pl	3sg2pl
qub	Huallaga Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Huallaga Huánuco Quechua	-manqui	-masha	-manqui				-manchi	-chimasha			-shunqui / -shcanqui		-shunqui	-mushcă		-shunqui
qvm	Margos-Yarowilca-Lauricocha Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union	-manqui		-manqui								-shunqui		-shunqui			
qxh	Panao Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union	-manqui	-masha				-mashcanqui					-shcanqui					
qxn	Northern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos	-manqui	-mushga		-manga			-mantsic						-yăshunqui			-yăshunqui
qxo	Southern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos	-manqui	-mushga		-manga			-mantsic						-yăshunqui			-yăshunki
qvh	Huamalíes-Dos de Mayo Huánaco Quechua	Quechua 1	Central Quechua 1	Huaylay		-munki / -manki	-man				-ykanki									-yăshayki	-yăshunki
qwh	Huaylas Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay		-huanqui	-huarca														
qvn	North Junín Quechua	Quechua 1	Central Quechua 1	Yaru Quechua		-manqui	-masha				-mashcanqui					-shcanqui					
qvw	Huaylla Wanca Quechua	Quechua 1	Central Quechua 1	Jauja-Huanca		-manqui	-man									-shunqui		-cushunqui	-chwanpis	-cushunqui	-cushunqui
quy	Ayacucho Quechua	Quechua 2	Ayacuchan Quechua			-wanki	-wan	-wankichik	-wanku		-wankiku	-wanku		-wanku	-ki	-sunki		-sunki	-ykichik	-sunkichik	
quh	Southern Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua	South Bolivian-Argentinian Quechua		-wanki	-wan		-wanku	-wanchej					-yki				-ykichej		
qul	North Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua			-wanki	-wan	-wankichis			-wankichis	-wanku		-wanku	-yki	-sunki			-ykichis		
quz	Cusco Quechua	Quechua 2	Cuscan Quechua			-wanki	-wan	-wankichis	-wanku	-wankichis	-wankiku	-wanku	-wankiku	-wanku	-yki	-sunki	-ykiku	-sunkiku	-ykichis	-sunkichis	-sunkichis
qve	Eastern Apurímac Quechua	Quechua 2	Cuscan Quechua			-wanki	-wan		-wanku		-wankikú				-yki				-ykichis		
qvc	Cajamarca Quechua	Cajamarca-Lambayeque Quechua	Cajamarca Quechua	Cajamarca Quechua		-wangi		-washpapis	-wananqa		-wangilla				-yki	-shushqa			-ykillapas	-shushpa	
quf	Lambayeque Quechua	Cajamarca-Lambayeque Quechua	Lambayeque Quechua	Lambayeque Quechua		-manki	-masha								-yki				-shaykillapa		
qxl	Salasaca Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua A	Tungurahua Highland Quichua		NONE				-gangui	NONE								-kiga		
qxr	Cañar Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Cañar-Azuay-South Chimborazo Highland Quichua		NONE	-manga	-huanguichic / -huarcani	NONE		-huangui				NONE	-huarca	NONE			-huanguichic	
qvi	Imbabura Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Imbabura Highland Quichua	-huangi			-jun	-chingui									-jpipash		
inb	Colombian Inga	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-wankungi / -wangi	-wanka				-rkangi	-murka			-iki				-ikichita		
qup	Southern Pastaza Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-wanki		-wankichi			-shkanki								-ptiypas		
quw	Tena Lowland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-huangi	-huaca / -shcami												-jpis	-shcami	-canchi
qvo	Napo Lowland Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-huanqui	-huarca														
qvz	Northern Pastaza Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-huangi	-hua		-huanaun										-cplga		
qvs	San Martín Quechua	San Martin-Amazonas Quechua	San Martín Quechua			-wanki	-washka				-shkanki				-ptinika						

Figure 4: Dialects are organized by their typological classification and with each clade in the family tree. Under the "verbs" row are columns of subject-object polypersonal verbal suffixation patterns.

ISO	Dialect	Family	Sub-family	Clade 1	Clade 2	Clade 3	Clade 4
qvc	Cajamarca Quechua	Cajamarca-Lambayeque Quechua	Cajamarca Quechua				
quf	Lambayeque Quechua	Cajamarca-Lambayeque Quechua	Lambayeque Quechua				
qxl	Salasaca Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua A	Tungurahua Highland Quichua			
qxr	Cañar Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Cañar-Azuay-South Chimborazo Highland Quichua			
qvo	Napo Lowland Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Napo Low-land Quechua
qup	Southern Pastaza Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Pastaza Quechua
quw	Tena Lowland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Pastaza Quechua
qvz	Northern Pastaza Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Pastaza Quechua
qvi	Imbabura Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Imbabura Highland Quichua		
inb	Colombian Inga	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia-Oriente Quechua	Colombia-Oriente Quechua		
qub	Huallaga Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Huallaga Huánuco Quechua		
qvm	Margos-Yarowilca-Lauricocha Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union		
qxh	Panao Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union		
qxn	Northern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos		
qxo	Southern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos		
qvh	Huamalíes-Dos de Mayo Huánuco Quechua	Quechua 1	Central Quechua 1	Huaylay			
qwh	Huaylas Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay			
qvw	Huaylla Wanca Quechua	Quechua 1	Central Quechua 1				
qvn	North Junín Quechua	Quechua 1	Central Quechua 1	Yaru Quechua			
quy	Ayacucho Quechua	Quechua 2	Ayacuchan Quechua				
qul	North Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua				
qve	Eastern Apurímac Quechua	Quechua 2	Cuscan Quechua	Cusco Quechua	Eastern Apurímac Quechua		
quz	Cusco Quechua	Quechua 2	Cuscan Quechua	Cusco Quechua			
quk	Chachapoyas Quechua	San Martin-Amazonas Quechua	Chachapoyas Quechua				
qvs	San Martín Quechua	San Martin-Amazonas Quechua	San Martín Quechua				

Table 11: ISO codes and their corresponding family information, including dialect classification, family, sub-family, and clade structure.

Language	ISO	Source
Huallaga Huánuco Quechua	qub	(Weber, 1989)
Cajamarca Quechua	qvc	(Quesada, 1976)
Colombia–Ecuador Quechua	qvi, qxr, qvo	(Carpenter, 1982)
Southern Conchucos Ancash Quechua	qxo	(Hintz, 2017)
Tena Lowland Quichua	quw	(Haboud de Ortega et al., 1982), (Orr, 1992), (Orr, 1973)
Northern Conchucos Ancash Quechua	qxn	(Wroughton, 1988)
Huaylla Wanca Quechua	qvw	(Raez, 1917), (Cordero Crespo, 1955)
Cañar–Azuay–South Chimborazo Highland Quichua	qxr	(Cerrón-Palomino, 1976), (Guzmán, 1920)

Table 12: Grammars with dialect-specific information on polypersonal verbal agreement.

C Appendix: Results

C.1 Neural Model Variants

In addition to the configurations reported in the main results, we conducted a series of exploratory experiments to better understand the contribution of linguistic features and input formatting in the neural model under the full data setting (see Table 13).

One consistent finding across these experiments is that the ordering of linguistic tags relative to the raw text has a sizeable effect on performance. When both polypersonal and lexical rules were included, an initial configuration appended tags after the text (text + tags), yielding slightly lower performance. Reversing this order (tags + text) resulted in consistent improvements across runs, with the strongest configuration achieving the highest overall performance of over 99% accuracy. Presenting linguistically enriched features at the beginning of the input sequence seems to guide the model more effectively by foregrounding relevant structural cues prior to processing the surface form.

A further refinement involved examining the contribution of individual affixes within the polypersonal rule set. In particular, the suffixes *man* and *ki* were hypothesized to introduce noise. The suffix *man* can encode non-polypersonal meanings such as directional or locative functions across dialects, while *ki* is both orthographically and phonologically short and appears in a wide range of nominal forms, potentially reducing its discriminative value. Empirical results, such as those in Table 13, supported this hypothesis: excluding these affixes led to a consistent improvement in performance. Consequently, the final neural configuration omits *man* and *ki* from the polypersonal feature set.

C.2 Naive Bayes Chunking Strategy

For the Naive Bayes classifier, we conducted additional experiments to evaluate the effect of document chunking on model performance (see Table 13). Specifically, we compared two configurations: one in which all available chunks from each document were used, and another in which the number of chunks per document was capped at 50.

The results show that limiting the number of chunks per document leads to a substantial improvement in performance. Without any restriction, documents with larger amounts of text contribute disproportionately more training instances, potentially biasing the classifier toward dialects with greater data volume. By capping the number of chunks per document, the dataset becomes more balanced across dialects, reducing this source of skew.

C.3 Detailed Segmentation Results

A detailed breakdown of segmentation performance by dialect family and data condition is shown in Table 7.

In high-resource settings (all data), differences between segmentation methods are relatively small. As shown in Table 7, performance remains consistently high across all segmentation strategies, with only marginal variation in both weighted and macro F1. In these conditions, standard frequency-based methods such as BPE and Unigram achieve the highest or near-highest scores across most families, which may indicate that sufficient data allows models to learn effective sub-word representations without explicit morphological guidance.

In contrast, segmentation choice has a slightly more pronounced effect in the low-resource setting. Across families, morphology-aware and hybrid approaches consistently outperform BPE when looking at macro F1. In particular, PRPE+BPE and Unigram yield the strongest and most stable performance. For example, in the Quechua I and Quechua II families, PRPE+BPE achieves the highest or tied-highest macro F1, while Unigram performs best for Quechua I under low-resource conditions. These results indicate that morphology-aware or proba-

Model	Data	Rules	Notes	Accuracy	Precision	Recall	F1
bayes	all	none	unlimited chunk size	.9145	.9062	.9145	.8971
bayes	all	none	limited chunks = 50	.9461	.9539	.9461	.9385
neural	all	verb+lexical	tags + text; no “man” or “ki”	.9953	.9953	.9953	.9951
neural	all	verb+lexical	text + tags	.9906	.9909	.9906	.9906
neural	all	verb	tags + text; all affixes	.9933	.9933	.9933	.9932
neural	all	verb	tags + text; no “man” or “ki”	.9940	.9940	.9940	.9940

Table 13: Additional experimental results on impact of chunk size constraints, rule-based linguistic features, and input ordering of tags versus text.

Setting	n	ΔF1	p (Wilcoxon)	d	F1 (rules)
All Data	27	+0.0015	0.133	0.324	0.957
Bible Only	25	+0.0008	0.285	0.148	0.996
No Bibles	10	+0.0649	0.398	0.342	0.570
NB x All (f)	27	+0.0241	0.502	0.234	0.534
B-only x NB	25	+0.0061	0.465	0.215	0.809

Table 14: Effect of linguistic rule augmentation across different data conditions. Settings using verb rules includes: all data, bible only, no bibles, and no bibles train by all data evaluation relaxed to family accuracy. Setting under lexical rules include: bible-only train by no bible evaluation. n denotes the number of dialects evaluated. ΔF1 indicates the average change in F1 score relative to the no-rules baseline. p (Wilcoxon) reports the significance of paired differences across dialects. d (Cohen’s d) measures effect size. F1 (rules) shows the macro F1 score of the rule-augmented model.

Dialect	Classifier Choice											Total
	qub	quf	quh	quk	qul	qup	quw	qux	quy	quz	other	
qub	**91**	0	0	0	0	0	0	0	0	1	1	93
quf	0	**121**	0	0	0	0	0	0	0	0	0	121
quh	0	0	**254**	0	0	0	0	1	0	9	0	264
quk	0	2	0	**0**	0	0	0	1	0	0	2	5
qul	0	0	0	0	**106**	0	0	0	0	0	0	106
qup	0	0	0	0	0	**133**	0	0	0	0	0	133
quw	0	0	0	0	0	0	**89**	0	0	0	0	89
qux	0	0	6	0	0	0	0	**144**	0	3	0	153
quy	0	0	0	0	0	0	0	0	**823**	1	0	824
quz	0	0	21	0	0	0	0	2	0	**2081**	1	2105
other	0	0	1	0	0	0	0	0	0	0	0	
Total	91	123	282	0	106	133	89	148	823	2095		

Table 15: Truncated confusion matrix for dialect classification on all data without rules.

bilistic segmentation may better support generalization when training data is limited.

Performance differences also vary by dialect family (see Figure 5 and Figure 6). The Quechua II model achieves the highest scores across nearly all conditions, reflecting its larger and more balanced training data. In contrast, the San Martín model shows consistently lower performance across segmentation strategies, suggesting greater sensitivity to data sparsity. However, even in these lower-resource families, PRPE+BPE improves macro F1 relative to BPE, which suggests that linguistically informed segmentation provides some gains.

Dialect	Classifier Choice											Total
	qub	quf	quh	quk	qul	qup	quw	qux	quy	quz	other	
qub	**91**	0	0	0	0	0	0	0	0	1	1	93
quf	0	**121**	0	0	0	0	0	0	0	0	0	121
quh	0	0	**247**	0	0	0	0	0	0	17	0	264
quk	0	1	0	**0**	0	0	0	1	0	0	3	5
qul	0	0	0	0	**106**	0	0	0	0	0	0	106
qup	0	0	0	0	0	**133**	0	0	0	0	0	133
quw	0	0	0	0	0	0	**89**	0	0	0	0	89
qux	0	0	4	0	0	0	0	**147**	0	2	0	153
quy	0	0	0	0	0	0	0	0	**823**	1	0	824
quz	0	0	1	0	0	0	0	0	0	**2104**	0	2105
other	0	0	1	0	0	0	0	1	0	0	–	
Total	91	122	253	0	106	133	89	149	823	2125		

Table 16: Truncated confusion matrix for dialect classification using verb rules on all data.

Dialect	Classifier Choice										Total
	quf	quh	qul	quy	quz	qve	qvo	qvw	qwh	other	
quf	**0**	0	0	0	0	0	0	0	0	0	0
quh	9	**8**	16	47	19	3	1	12	11	35	161
qul	0	0	**0**	0	0	0	0	0	0	0	0
quy	0	0	0	**394**	0	0	0	1	1	0	396
quz	165	3	24	93	**1020**	198	14	92	24	38	1671
qve	0	0	0	0	0	**0**	0	0	0	0	0
qvo	0	0	0	0	0	0	**0**	0	0	14	14
qvw	0	0	0	0	0	0	0	**0**	0	0	0
qwh	0	0	0	2	0	0	0	1	**4**	7	14
other	0	0	0	0	1	0	0	0	0	–	
Total	174	11	40	536	1040	201	15	106	40		

Table 17: Truncated confusion matrix for dialect classification on no rules model trained on only-Bible data tested on no-Bible data.

Dialect	Classifier Choice										Total
	quf	quh	qul	quy	quz	qve	qvo	qvw	qwh	other	
quf	**0**	0	0	0	0	0	0	0	0	0	0
quh	8	**7**	17	50	25	1	3	3	21	26	161
qul	0	0	**0**	0	0	0	0	0	0	0	0
quy	0	0	0	**394**	0	0	0	0	1	1	396
quz	35	8	29	97	**1359**	49	17	18	33	26	1671
qve	0	0	0	0	0	**0**	0	0	0	0	0
qvo	0	0	0	0	0	0	**0**	0	0	14	14
qvw	0	0	0	0	0	0	0	**0**	0	0	0
qwh	0	0	0	3	0	0	0	0	**7**	5	15
other	0	0	0	0	1	0	0	0	0	–	
Total	43	15	46	544	1385	50	20	21	62		

Table 18: Truncated confusion matrix for dialect classification on lexical model trained on only-Bible data tested on no-Bible data.

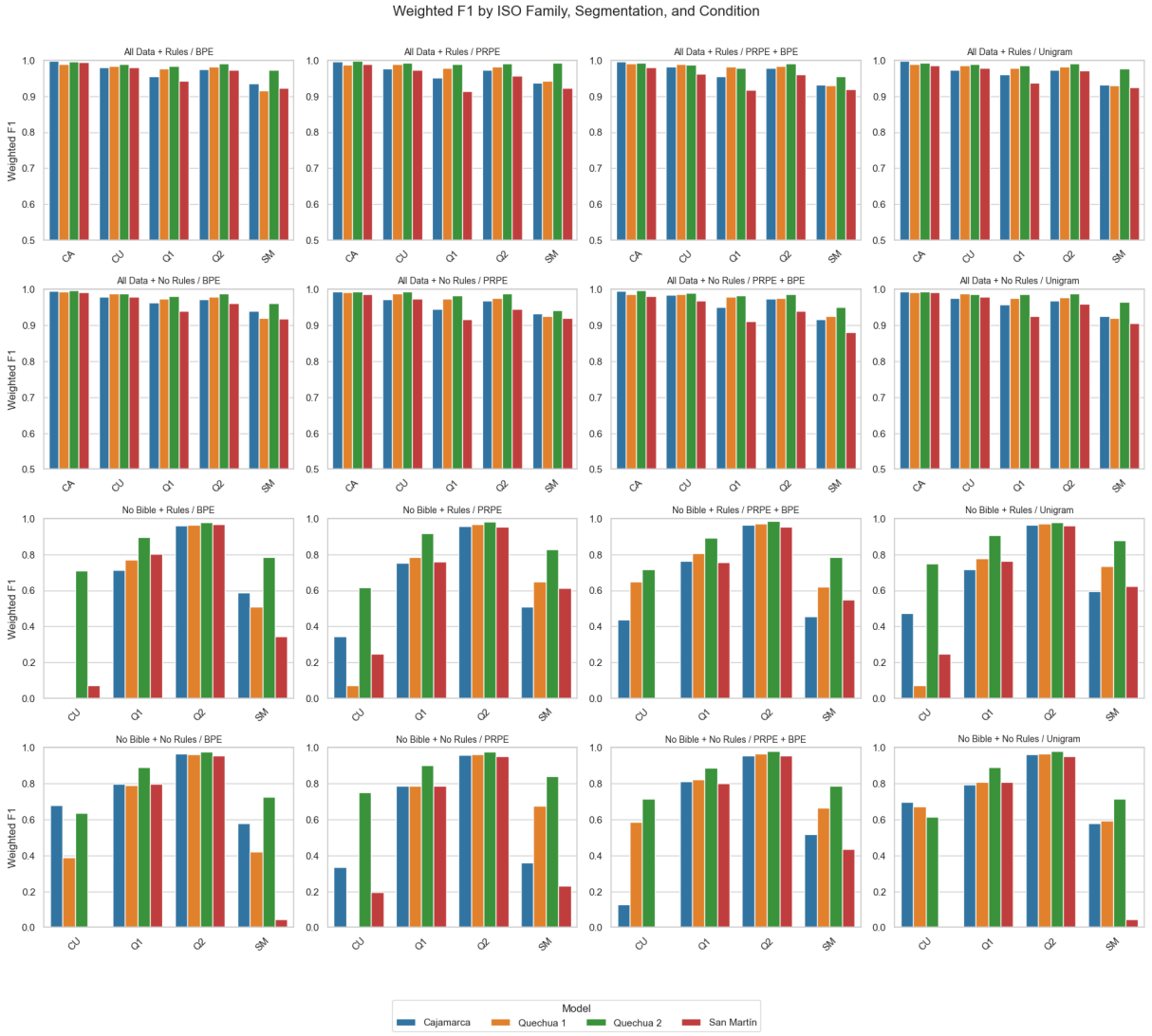

Figure 5: Weighted F1 scores by language family, segmentation method, and experimental condition. Each subplot is a combination of data and rules; bars show performance trained on different Quechua families.

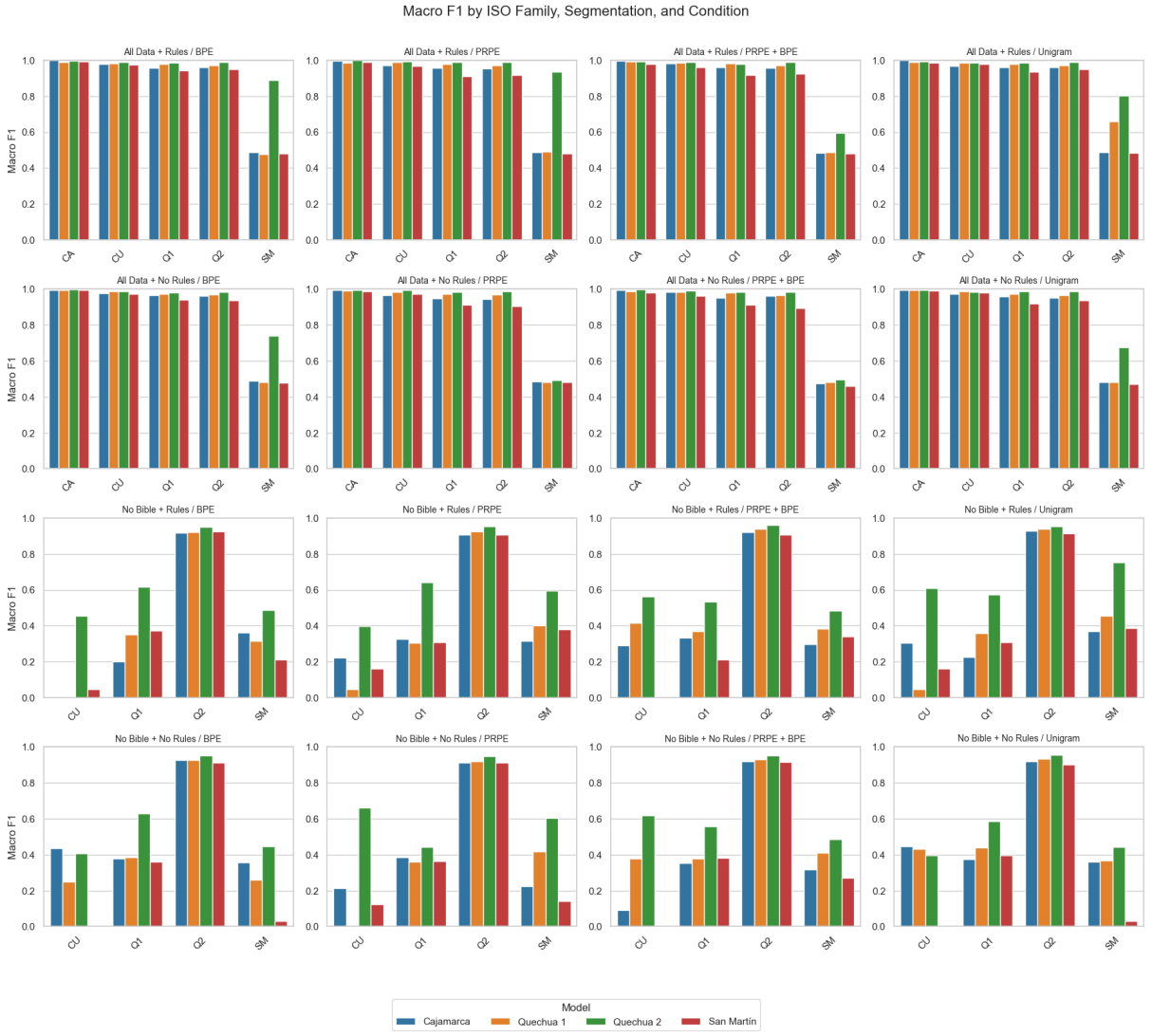

Figure 6: Macro F1 scores by ISO language family, segmentation method, and experimental condition. Each subplot is a combination of data and rules; bars show performance trained on different Quechua families.