Morphological Parsing for Media Lengua: When Accessibility Matters More Than State-of-the-Art

Jesse Stewart and Olga Kriukova
University of Saskatchewan
stewart.jesse@usask.ca, olga.kriukova@usask.ca

Abstract

While machine learning approaches dominate contemporary NLP research (Vylomova et al., 2020), a critical gap exists between published models and tools actually used by target communities (Gessler and von der Wense, 2024). This paper presents two morphological parsers for Media Lengua (ISO 639-3: mue), an endangered mixed language of Ecuador, demonstrating that a JavaScript rule-based system (98.6% accuracy) can outperform a CRF model (95.7% F1) while offering immediate community accessibility.

Not all language structures permit straightforward rule-based parsing; however, when a language's morphology allows for this approach with competitive accuracy (cf. Vylomova et al., 2020), we argue that it should be preferred for its practical advantages: immediate browser-based deployment, transparency, zero infrastructure requirements, and long-term maintainability. Our rule-based parser runs entirely in the browser, is freely available online, and can be adapted to other Quechuan languages. In contrast, while the CRF model performs well on benchmarks, it requires additional infrastructure to become accessible.

Our comparison highlights the need to evaluate NLP tools not only on accuracy metrics but also on accessibility and real-world adoption, which is particularly crucial for endangered language communities where sustainable, community-accessible tools can support language documentation, education, and revitalization.

1 Introduction

The development of natural language processing tools for endangered languages faces a critical challenge: while computational models continue to advance in performance on benchmark tasks, a significant gap persists between published models and tools actually accessible to the communities they are meant to serve (Gessler and von der Wense, 2024). Morphological processing tools in particular have significant potential to aid language documentation efforts for endangered languages (Wiemerslage et al., 2022), yet this potential remains unrealized when tools require technical infrastructure that communities lack.

In this study, we address this problem through the development of morphological parsing tools for Media Lengua, an endangered mixed language which has Spanish-origin vocabulary and Quichua-origin morphosyntax. Media Lengua is spoken by approximately 1,204 people in communities near Lago San Pablo, Imbabura, Ecuador and by approximately 1,703 people in southern Cotopaxi, Ecuador (Stewart et al., 2023). The language was formed primarily through the process of relexification, replacing an estimated 90% of native Quichua words with Spanish-origin words (Muysken, 1981, 1997). Table 1 below exemplifies Media Lengua structure: all roots are of Spanish origin (bolded) and all grammatical morphemes are of Quichua origin. As a mixed language, Media Lengua emerged not from communicative necessity but for expressive purposes among proficient bilinguals (Meakins and Stewart, 2022), resulting in relatively regular agglutinative morphology with systematic divisions between elements from each source language (Meakins, 2013; Meakins and Stewart, 2022).

Orth	*Mio hijapash Quitopi.*		
Parse:	**mio**	**ixa**-paʃ	**kito**-pi
Sp:	mi	hija-CONJ	Quito-LOC
Q:	ñucapa	ushi-CONJ	Quitu-LOC
En:	my	daughter-CONJ	Quito-LOC
Trans:	My daughter is in Quito as well.		

Table 1: Media Lengua parsing example

This morphological regularity presents an opportunity to examine a fundamental question in NLP tool development: when a language's structure per-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 1–9

mits accurate rule-based parsing–and evidence suggests rule-based systems remain competitive or superior in very low-resource settings (Vylomova et al., 2020)–should we default to machine learning approaches, or prioritize methods that offer immediate community accessibility? We introduce two morphological parsers for Media Lengua that embody different answers to this question. The first is a rule-based morphosyntactic parser (hereafter RB-parser) achieving 98.6% accuracy and designed for immediate community access through a browser-based interface. The second is a morphosyntactic parser based on a classic Conditional Random Fields (CRF) model (hereafter CRF-parser), achieving an F1-score of 95.7% and intended primarily for linguists to facilitate parsing of texts.

Our comparison demonstrates that for endangered language communities, the RB-parser's combination of high accuracy and zero-barrier deployment provides greater practical value than approaches that require backend infrastructure and technical expertise to operate, despite their advantages in training efficiency. We argue that accessibility and long-term sustainability—factors rarely measured in computational linguistics research—are essential considerations for tools meant to support language documentation and revitalization.

1.1 Design philosophy and applicability

Both the RB-parser and CRF-parser share the same fundamental objective: accurate morphological segmentation of Media Lengua. However, they differ significantly in their design philosophy and the contexts in which they are most applicable, reflecting distinct cases when NLP tools can serve endangered language communities.

The RB-parser is designed for maximum accessibility to speakers, learners, and linguists without technical barriers. The parser performs segmentation and broad IPA transcription to reflect general pronunciation patterns, with each grammatical morpheme distinctly separated by dashes, ensuring that the output aligns closely with the parse tier illustrated in Table 1. In addition, the parser provides interlinear glosses offering approximate translations of the lemmas in Media Lengua's source languages—Quichua and Spanish—as well as in English, along with standard glossing abbreviations for grammatical morphemes. The development of this parser was facilitated by existing lexicographical resources: comprehensive verb and non-verb dictionaries compiled from prior fieldwork, along with documented inventories of grammatical morphemes (see Stewart et al., 2020). This allowed development efforts to focus on implementing parsing logic rather than resource creation.

To ensure community accessibility, the parser has been developed as a browser-based application requiring no installation, server access, or technical expertise. It runs entirely in the user's browser using JavaScript, with a user-friendly interface designed through HTML and CSS. This choice of platform facilitates ease of use for various stakeholders, since most households in Media Lengua communities have Internet access through PCs and/or smartphones. The parser is hosted online[1] and available for free, where it can be used immediately by anyone with a web browser.

The parser's design was shaped by ongoing community-facing work with Media Lengua speakers and consultants, including review of dictionary entries, discussion of potential uses, and informal feedback on parser outputs, though it has not been evaluated through a formal user study and we therefore avoid making strong claims about measured usability or adoption. Community use is expected to centre on language learning, text preparation, checking morphological segmentation, and supporting local documentation efforts. Because the tool is browser-based and the dictionaries are transparent, community feedback can be incorporated through concrete corrections to lexical entries, orthographic variants, and morpheme analyses rather than requiring model retraining.

By contrast, the CRF-parser is applicable to the contexts where linguists need to process larger volumes of new language data efficiently. Importantly, its prospective users are limited to linguists and community members who already have some level of computational expertise.

The CRF approach was selected for its suitability to agglutinative morphology and practical advantages in low-resource contexts (Ruokolainen et al., 2013). Additionally, CRF models require no specialized hardware and produce interpretable models that facilitate error analysis. The resulting models are compact, fast, and have minimal software dependencies, supporting long-term reproducibility. While not state-of-the-art, CRF models provide a stable, well-understood baseline for

[1] http://jessestewart-ling.github.io/languagetools/Parser.html

comparison with rule-based methods without the complexity and resource requirements of neural architectures (cf. Kriukova et al., 2025; Wiemerslage et al., 2022).

The CRF-parser has potential to reduce the the so-called "annotation bottleneck" (Foley et al., 2018; Moeller, 2021) by accelerating preliminary morphological segmentation. However, even though it performs well on benchmark metrics, using this model in practice requires setting up a server, creating a web interface, and maintaining the infrastructure—barriers that often prevent tools from reaching community members or linguists without computational background who need them. The model is made available on GitHub[2] for linguists who work with Media Lengua and other Quechuan languages alike.

2 Media Lengua Structure

Media Lengua is an agglutinating language with SOV word order. Its grammatical morphology is highly regular and can be categorized as verbal and non-verbal (nouns, adjectives, adverbs etc.) with clitics that can attach to both. Grammatical morphemes uniquely suffix to roots and can build in complexity extending to the right. The lack of grammatical irregularities and predictable morphonology in Media Lengua make it an ideal test case for the type of parser described in this paper. Yet one primary challenge facing the parser is the lack of a standard orthography (cf. Rios Gonzales and Castro Mamani, 2014), which makes user input variable (e.g., the word *daiy* 'from there' has at least 15 documented spelling variations (Stewart et al., 2020)).

3 RB-parser

3.1 Data

Data for this parser comes from the only published Media Lengua dictionary (Stewart et al., 2020), which contains 3210 lemmas and 1974 orthographic variations. These lemmas are complemented by 24 hours of glossed conversational, narrative and elicited speech data housed at The Archive of the Indigenous Languages of Latin America (Stewart and Prado Ayala, 2025), which provide additional instances of orthographic variation and morpheme cluster combinations.

[2] https://github.com/HelgaKr/ML_CRF

3.2 Parser design

This rule-based parser leverages Media Lengua's regular morphological structure through dictionary lookup and pattern matching. This design choice prioritizes immediate implementation and transparency over statistical complexity. The parser employs left-to-right processing–a well-established approach for agglutinative languages (Jarzabek and Krawczyk, 1975; Weber, 1989)–with an implementation optimized for JavaScript execution in browser environments.

The parser operates with five JSON dictionaries containing 9,965 predefined entries, optimized for real-time processing. Two dictionaries contain grammatical morphology: verbal morphology (178 entries) and non-verbal morphology (113 entries). Both dictionaries include documented orthographic variations and plausible, yet not documented, spelling variants. Additionally, they contain a set of clitics that can attach to both verbal and non-verbal morphology. The structure of these dictionaries is identical and includes an orthographic and IPA representations of the morpheme, and the morpheme's gloss.

The largest JSON dictionaries contain verb (1,848) and non-verb (6,847) entries, reflecting Media Lengua's noun-heavy lexicon (Stewart et al., 2020). Both dictionaries also contain documented orthographic variations, including common typos, and other plausible variations, though this coverage is not exhaustive. The structure of these dictionaries is similar though not identical. Both contain orthographic and IPA representation of the lemma, broad translations (Quichua, Spanish, & English), and the origin of the lemma (Spanish or Quichua). Additionally, each verb object contains an IPA representation of the root form of the verb (i.e., with infinitive morphology removed: *comina* /komina/ 'eat' -> /komi-/).

The fifth dictionary contains 979 morphemes and morpheme clusters (e.g., *-gucunata* /-gu-kuna-ta/ '-DIM-PL-ACC') extracted from 12 hours of speech data. These clusters are exclusively used in the second parsing algorithm (see 3.6). This dictionary contains only IPA-representations of morphemes and morpheme clusters.

The parser uses a two-stage approach (see Sections 3.5 and 3.6): a primary parsing algorithm attempts direct dictionary matching against existing entries. If this attempt fails, a secondary predictive algorithm segments roots from grammatical

morphemes for lemmas not found in the verb and non-verb JSON dictionaries.

3.3 IPA conversion

Before user input enters the algorithms, it undergoes preprocessing to normalize orthographic variation. The input is lowercased, converted to IPA, and stripped of all punctuation using 94 regular expressions. These regular expressions are specifically ordered to convert a word one phoneme or phoneme cluster at a time resulting in an accurate phonemic IPA representation based on the Media Lengua phonotactics. Multi-character sequences are processed before their constituent parts to ensure correct phoneme identification. For example, <lll> in *iguallla* is converted to /lʒ/ (resulting in /igualʒa/), before the <ll> -> /ʒ/ rule applies, to avoid the incorrect result */iguaʒla/.

This normalization substantially reduces orthographic variation and thus the number of entries required in JSON dictionaries. For example, variants of *acienda* /asienda/ 'ranch' (*hacienda, asienda, hasienda*) can be reduced to one entry (/asienda/). The user input is then converted from a string to an array using space as the delimiter to capture each individual word. In cases when user input contains typos, the parser defaults to the predictive algorithm.

3.4 Compounds

Media Lengua contains numerous compound words that are often written as separate tokens (e.g., *choclo tanda* 'cornbread'). To prevent the parser from analyzing only the first component of such compounds (e.g., *choclo* 'maize'), the system implements a compound detection algorithm. The parser tests for compounds by appending the first two segments of the following word to the current word (e.g., *choclo ta*) and matching this combination as a regular expression pattern against the dictionaries. Matched compounds are treated as single units in subsequent processing. This function iterates to detect compounds containing up to four words. Further details and limitations of this approach are discussed in Section 5.

3.5 Parsing algorithm

The parsing algorithm identifies lemmas through incremental substring matching against lexical dictionaries. Beginning with the first character of the input, progressively longer prefixes are generated up to 22 characters (exceeding any dictionary entry length). For the input *perromanmi* /pezo̱manmi/ 'dog-DIR-VAL', the algorithm generates prefixes of increasing length: 'p', 'pe', 'pez̨', continuing through the complete nine-character string 'pezo̱manmi' (see Figure 1).

Figure 1: A word split into incremental substrings

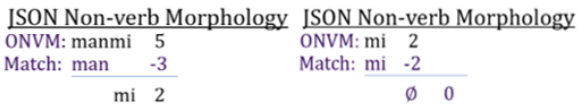

Figure 2: Segmentation of grammatical morphemes

Each prefix is compared independently against both verb and non-verb lexical dictionaries. The algorithm selects the longest matching prefix from each dictionary. In our example, the four-character prefix 'pezo̱' matches a non-verb entry meaning 'dog', while no verb match is found. Translations in the source languages (Quichua *alcu*, Spanish *perro*, and English *dog*) are extracted from the matched entry for the final output.

The remaining unmatched portion of the input becomes the candidate for morphological segmentation. By subtracting the matched lemma length from the total input length, the algorithm isolates potential grammatical morphemes. In our example, removing the four-character lemma *pezo̱* from the nine-character input leaves 'manmi' as the morphological material for the non-verb parse. The verb parse, having found no lemma match, retains the entire input "pezo̱manmi" for morpheme analysis.

This process is applied recursively to segment grammatical morphemes. Progressively longer prefixes (up to eight characters, the maximum documented morpheme length) are generated from the candidate string and compared against morphology dictionaries. For 'manmi', both a two-character match ('ma') and three-character match ('man') are found. The algorithm selects the longer match 'man' (directional marker, glossed as DIR), leaving 'mi' for further analysis. In the next iteration, 'mi' matches the validator marker (VAL). This recursive segmentation continues for up to ten iterations, accommodating the maximum documented morpheme count in Media Lengua.

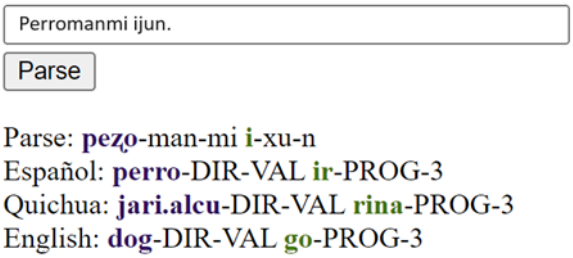

Figure 3: RB-based parser output

After completing all iterations, both the verb and non-verb analyses are reconstructed and compared against the original input. The verb analysis, having failed to identify a lemma, produces an unsegmented result, while the non-verb analysis successfully reconstructs 'pezomanmi' through the sequence *pezo-man-mi*. This successful reconstruction confirms the non-verb classification, enabling part-of-speech-based visual coding in the interface to assist learners. As shown on Figure 3, the extracted glosses and translations are then formatted as interlinear output, and processing advances to the next word.

When neither analysis successfully reconstructs the input, the system invokes the predictive algorithm (see Section 3.6) to attempt partial segmentation. The complete algorithmic specifications and implementation details are available in the project GitHub repository[3].

3.6 Predictive algorithm

The predictive algorithm addresses inputs absent from the lexical dictionaries by attempting to identify suffix boundaries using corpus-extracted morpheme clusters. Unlike the parsing algorithm's left-to-right approach, this method works from right to left, progressively testing shorter suffixes against a dictionary of 979 attested morpheme clusters derived from 12 hours of transcribed speech.

The algorithm extracts suffixes of decreasing length from the input, beginning with the final 15 characters (the maximum observed cluster length). For *perromanmi* /pezomanmi/, which contains only nine characters, the algorithm initially considers the entire word as a potential suffix. It then systematically removes characters from the left: first testing 'ezomanmi', then 'zomanmi', then 'omanmi', continuing through single-character suffixes.

[3] https://github.com/JesseStewart-LING/languagetools

Figure 4: Parsing result with an unrecognized cluster

Each candidate suffix is matched against the morpheme cluster dictionary, proceeding from longest to shortest to maximize the identified suffix material. In this example, the five-character sequence 'manmi' matches a documented cluster. Subtracting this suffix length from the input yields 'pezo' as the predicted root, with the cluster's gloss (-man-mi 'DIR-VAL') extracted for display.

The algorithm does not attempt further decomposition of matched clusters into individual morphemes. This design reflects a fundamental limitation: without lexical verification of the predicted root, part-of-speech assignment remains uncertain, and morpheme selection rules critically depend on POS distinctions (e.g., verbal versus nominal paradigms). Nevertheless, identifying the root-suffix boundary provides valuable segmentation information even without complete morpheme-by-morpheme analysis.

Given the incomplete coverage of the morpheme cluster dictionary—which captures frequent but not exhaustive combinations—predicted segmentations are visually distinguished (displayed in red) and accompanied by an explicit warning in Spanish. Users are informed that analyses in red represent approximations and are invited to report disagreements via email, enabling iterative refinement of the system through community feedback (see Figure 4).

4 CRF-parser

4.1 Training data

The CRF model was trained on morphologically segmented data from the Media Lengua corpus collected by Kriukova. Each entry was converted to broad IPA transcription to reduce spelling variation (3.3). Table 2 shows the train-development-test split. Due to ongoing community consultation, the corpus cannot be made publicly available at this

time.

Split	Word types
Training	399
Development	199
Testing	200

Table 2: Word types distribution

4.2 CRF-parser design

Conditional Random Fields (CRFs) are probabilistic models that predict morpheme sequences by modeling transition probabilities between morphemes as a function of input features (Lafferty et al., 2001; Ruokolainen et al., 2013). We implemented our CRF using CRFsuite (sklearn-crfsuite) with the averaged perceptron algorithm (Collins, 2002). After hyperparameter optimization, the final model was trained with ð = 8 and maximum iterations = 150. The relatively low iteration count reflects the small dataset size.

Following Ruokolainen et al. 2013, we used four segmentation categories: B (beginning), M (middle), and E (end) for multi-character morphemes, and S for single-character morphemes. For example, *casakunamanta* 'house-PL-ABL' is labeled:

k	a	s	a	k	u	n	a	m	a	n	t	a
B	M	M	E	B	M	M	E	B	M	M	M	E

The model used character-level features to capture local orthographic and phonological context after IPA conversion, including the current character, neighbouring characters within the selected window, and boundary-position information. Evaluation was conducted against morpheme-boundary labels in the test set using these B/M/E/S categories.

4.3 Performance

The CRF-parser achieved an F1-score of 95.7% (see Table 3), demonstrating that statistical models can perform well even with limited training data (399 word types). By comparison, the rule-based parser achieved 98.6% accuracy on the same test set (97.5% when evaluated on the CRF test data specifically).

Precision	96.8
Recall	94.7
F1-score	95.7

Table 3: Model testing results

The CRF-parser's errors primarily stem from lack of part-of-speech information. For instance, it incorrectly parses *agua-man-ka* 'water-DIR-TOP' as *agua-ma-nka*, attaching the verbal morpheme *-nga* (3SG.FUT.UNCERT) to a noun. The rule-based parser avoids such errors through dictionary lookup that includes POS tags. Similarly, the CRF-parser occasionally fails to segment valid lemmas, treating *yu-ca* 'I-TOP' as a single unsegmented form *yuca* (also a lexical item 'cassava root').

While the CRF-parser demonstrates competitive performance and may offer convenience if large portions of data require annotation, its deployment for community use would, again, require server infrastructure and ongoing maintenance—barriers absent from the browser-based rule-based parser. The model is available at GitHub[4] for researchers working with Media Lengua and related Quechuan languages.

5 Conclusion

This paper has presented two morphological parsers for Media Lengua, demonstrating that the choice between rule-based and machine learning approaches should be guided not only by benchmark performance but also by accessibility, deployability, and the practical needs of target users.

The rule-based parser achieves 98.6% accuracy while offering immediate community access through a browser-based interface requiring no technical infrastructure (only HTML and CSS interface). It provides morphological segmentation alongside translations in Quichua, Spanish, and English, making it particularly valuable for language learners and speakers. The parser has been extensively tested using the Media Lengua corpus and by native speakers with positive feedback. Importantly, it can be adapted to other varieties: the Cotopaxi dialect requires only minor modifications to IPA conversions and lexical entries, while Quichua varieties can be supported by replacing the lexical dictionaries, given the shared grammatical structure.

The CRF-parser achieves an F1-score of 95.7% despite being trained on only 399 word types, demonstrating that statistical approaches can succeed with limited data for morphologically regular languages. The strong performance of the model is partly attributable to Media Lengua's concatenative, agglutinative structure—a characteristic

[4] https://github.com/HelgaKr/ML_CR

it shares with Quichua and other Quechuan languages. This suggests that that similar models could be developed for related low-resource languages. However, deploying this model for community use would require server infrastructure and ongoing technical maintenance, which would limit its accessibility compared to the rule-based browser alternative.

Our comparison highlights a critical gap in NLP research: while the field prioritizes benchmark performance, factors like accessibility, transparency, and long-term sustainability—essential for endangered language communities—receive far less attention. For Media Lengua, an immediately usable tool provides greater practical value than a model that requires computational expertise.

We argue that when a language's structure permits accurate rule-based parsing, such approaches should be seriously considered alongside machine learning alternatives, particularly for endangered language applications. The rule-based parser not only serves current documentation and learning needs but also enables community members to understand and modify linguistic rules directly, fostering local involvement in language technology development. This is especially important for revitalization efforts, where sustainability and community engagement are paramount (Bird, 2020; Czaykowska-Higgins, 2009).

Although our study does not introduce novel algorithmic approaches, it demonstrates that thoughtful tool design—prioritizing accessibility over cutting-edge methods—can have greater real-world impact. As the NLP community continues to develop tools for low-resource and endangered languages, we encourage researchers to evaluate their work not only on accuracy metrics but also on whether target communities can actually access and use these tools. The success of language technology should ultimately be measured by adoption and utility, not just performance on test sets.

Our primary future directions involve investigating hybrid architectures that combine rule-based parsing for high-confidence dictionary matches with statistical models for novel or borrowed forms, while preserving accessible browser-based deployment. Moreover, we plan to add a function that will show the parsing alternatives for words that may be parsed differently depending on the context (e.g., *yu-ca* 'I-TOP' vs. *yuca* 'cassava root'). We are also considering adaptation of the rule-based parser to related Quechuan languages, many of which face similar challenges: limited infrastructure, small datasets, and endangered status. The present parser has high potential for adaptability to related languages, provided they have at least partial documentation of vocabulary and morphosyntactic rules.

Limitations

While the rule-based parser achieves 98.6% accuracy, its performance heavily depends on dictionary coverage. This trade-off exemplifies the broader tension between rule-based and machine learning methods: the former requires explicit documentation of linguistic knowledge but offers transparency and modifiability, while the latter can generalize from patterns but remains opaque to non-specialists attempting to understand or correct errors.

The majority of parsing errors (1.4%) are due to morphological homography, where identical forms serve multiple grammatical functions (e.g., *-ta* functions as both accusative and interrogative marker). To date, 123 such instances have been identified and resolved using context-sensitive rules, reducing the error rate to less than 0.5%. The remaining errors are primarily false positives resulting from partial matches. For example, *estaca* /estaka/ 'stake' lacks a dictionary entry, but *esta* /esta/ 'this' and *-ca* /ka/ 'TOP' both exist, producing the incorrect parse *esta-ca* 'this-TOP'.

Moreover, detection of compound words presents some algorithmic challenges that affect both precision and recall. The parser tests multiword combinations by examining the first two characters of the following word—a compromise designed to balance competing error types. Testing only one character risks false positives: *Choclo tomay!* /tʃoklo tomai/ 'take the corn' might be incorrectly identified as a compound. Testing the entire following word risks false negatives when grammatical morphemes prevent dictionary matches. Testing two characters accommodates common Spanish-origin function words (*de* 'of', *al* 'to the') that appear in compounds; however, false positives remain possible when coincidental orthographic overlap occurs (e.g., *Choclo talvez?* "Maybe corn?" shares the initial sequence /tʃoklo ta-/ with the compound *choclo tanda* 'cornbread').

Additionally, the IPA transcription accuracy depends on input conforming to Media Lengua phonotactics. Borrowings from languages other than Spanish or Quichua (e.g., English *check* with <ck>)

may be transcribed incorrectly (e.g., */kk/).

These limitations in some way highlight an important advantage of the rule-based approach for community-based language work: errors are transparent, debuggable, and fixable without specialized expertise. Users who encounter errors can report specific cases, and additions to the dictionaries can immediately improve performance for all users. Such feedback loop that supports iterative, community-driven improvement. By contrast, addressing errors in the CRF-parser would require collecting additional training data, model retraining, and redeployment—processes that create barriers to community participation in tool refinement.

The rule-based parser's reliance on explicit dictionary entries reflects a deliberate design choice prioritizing accessibility over the ability to parse unseen forms. The absence of neural or transformer-based baselines reflects the same logic. Such models are now central to NLP and are not irrelevant to morphological segmentation; however, they generally require larger datasets, more technical infrastructure, and deployment conditions that are not aligned with the primary goal of this project. Given that the CRF model itself was trained on only 399 word types, we treat it as a lightweight statistical comparison rather than a comprehensive benchmark against the current state of the art. For endangered language communities, this trade-off favours long-term usability over state-of-the-art performance, and future work should evaluate neural and hybrid architectures once larger annotated datasets and appropriate deployment pathways become available.

Ethical Considerations

To make this parser as accurate as possible, we relied on the knowledge of native speakers of Media Lengua, Quichua, and Spanish who reviewed all the entries in the dictionaries used by the parser. They were adequately monetarily compensated for their time as per the research ethics board approval BEH 16-151 granted by the University of Saskatchewan. Additionally, as per our REB approval, we discussed any potential risks with the annotators, how the data would be used, and consent forms were signed. This parser was specifically designed for speakers, learners, and linguists interested in better understanding the morphological structure of Media Lengua. As such, it is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License and made freely available at GitHub[5], as well as the model[6].

Acknowledgments

We would like to thank Lucia Gonza Inlago, Gabriela Prado Ayala, and Mahyli Calapi for reviewing the JSON entries used in the parser. Your knowledge and help have substantially improved its accuracy. The research for this project was partially funded by SSHRC IDG 430-2018-00032.

References

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8. Association for Computational Linguistics.

Ewa Czaykowska-Higgins. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within canadian indigenous communities.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, Mark T. Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (Elpis). In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209.

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–6, Mexico City, Mexico. Association for Computational Linguistics.

Stanislaw Jarzabek and Tomasz Krawczyk. 1975. Ll-regular grammars. *Information Processing Letters*, 4(2):31–37.

[5] https://www.jessestewart.net/languagetools/Parser.html, https://github.com/JesseStewart-LING/languagetools

[6] https://github.com/HelgaKr/ML_CRF

Olga Kriukova, Katherine Schmirler, Sarah Moeller, Olga Lovick, Inge Genee, Antti Arppe, and Alexandra Smith. 2025. AI for interlinearization and POS-tagging: Teaching linguists to fish. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 139–149, Honolulu, Hawaii, USA. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Felicity Meakins. 2013. *Mixed Languages*, pages 159–228. De Gruyter Mouton, Berlin, Boston.

Felicity Meakins and Jesse Stewart. 2022. *Mixed Languages*, page 310–343. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Sarah Moeller. 2021. *Integrating machine learning into language documentation and description*. Phd thesis, University of Colorado.

Pieter C Muysken. 1981. *Halfway between Quechua and Spanish: The case for relexification*, pages 52–78. Karoma Publishers.

Pieter C Muysken. 1997. *Media Lengua*, pages 365–426. Karoma Publishers.

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.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria. Association for Computational Linguistics.

Jesse Stewart, Lucia Gonza Inlago, and Gabriela Prado Ayala. 2023. Cotopaxi media lengua is still very much alive. *Language Documentation Conservation*, 17:49–63.

Jesse Stewart and Gabriela Prado Ayala. 2025. Media lengua collection of Jesse Stewart. The Archive of the Indigenous Languages of Latin America (AILLA): 1923.

Jesse Stewart, Gabriela Prado Ayala, and Lucia Gonza Inlago. 2020. *Media Lengua dictionary*. Dictionaria.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, and 9 others. 2020. SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

David J Weber. 1989. A morphological parser for linguistic exploration. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session*, 33.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological Processing of Low-Resource Languages: Where We Are and What's Next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.