Authors: Olga Kriukova, Jesse Stewart
Affiliation: University of Saskatchewan
Title: Media Lengua and Imbabura Quichua LL Morphosyntactic Parser

This paper presents technology demonstration of a left-to-right morphosyntactic parser (LL-parser) of Media Lengua and Imbabura Quichua (ISO 639-3: QVI). Media Lengua (ISO 639-3: MUE) is an endangered mixed language with a Spanish-origin vocabulary and Quichua-origin morphosyntax, spoken by approximately 2,000 people in communities near Lago San Pablo, Imbabura, Ecuador (0.1727° N, 78.1921° W). Media Lengua formed primarily through the process of relexification, replacing an estimated 90% of the native Quichua Spanish-origin words. See example (1) where all roots are of Spanish origin and all suffixation is of Quichua origin. Both Media Lengua and Quichua are highly agglutinating SOV languages that make use of highly regular suffixation and verbal inflections.

(1)

Orth	*Ellami ponin vacunata aretesta ya haymi pasan centro agricolaman.*								
Parse:	eʒa-mi	poni-n	bakuna-ta	aretes-ta	ja	ai-mi	pasa-n	sentro	agrikola-man
Sp:	ella-VAL	poner -3	vacona-ACC	marca auricular-ACC	ya	de ahí -VAL	pasar-3	centro	agrícola-VAL
Q:	pay-VAL	churrana-3	vacona-ACC	zarcillus-ACC	ña	chaymanda-VAL	pasana-3	centro	agricula -VAL
En:	she-VAL	put -3	heifer-ACC	ear tag-ACC	then	from there -VAL	go to-3	hub	agriculture-VAL
Trans:	After she tags the calf, it heads to the agricultural center (not provided by the parser).								

**Parser output: Verbs in red.*

Parser input was captured and converted to IPA using extensive regular expressions, which take into count a wide range of spelling variations typical of both languages. Additionally, the parser relies on the Media Lengua dictionary (Stewart, Prado Ayala & Gonza Inago, 2020) and a limited number of entries from the *Kichwa-English-Spanish Dictionary* written by Kinti-Moss & Masaquiza Chango (2018). Both sources are stored in the parser script as databases which are used to identify a roots. Extensive entries were added to the databases to cover additional variants in spelling. The databases contain information such as Spanish, Quichua, and English translations, part of speech, IPA transcriptions, and spelling variations. In order to identify morphemes, the parser relies seperate databases of Media Lengua and Quichua verbal and nominal morphemes, which also contain extensive variants to capture a wide range of possible input.

The parser is written in JavaScript and tailored to work with Media Lengua and Quichua though a number of diagnostics including a matrix-based a approach to match lemmas and morphemes from the database. The script takes a Media Lengua or Quichua utterance of any size (from one word to an entire text). The output not only provides a parsing result but also a translation of each root and gloss of each morpheme. Additionally, the script automatically compiles a list of gloss abbreviations and their meanings for the parsing result for user convenience.

At the time of writing, the accuracy rate of the parser is estimated at 97%, barring typos and words not present in the dictionaries. We are looking add a prediction function parse words not found in the datasets.

Keywords: Media Lengua, Quichua, LL-parser, morphosyntactic parser, JavaScript.