Quechua Multi-Varietal Text Classification

Intro

Related Work

Prompt

Submissions should be in the ACL format. The proposal should be **no longer than 2 pages (minus references)** and include the following sections:

1. **Introduction Section**, which should explain the context of the project. It should contain the following information:
   - *Task / Research Question Description*: What is the task you are trying to solve or what is the research question you are trying to answer? What model/method are you proposing to use for this task?
   - *Motivation & existing work*: Why should we care about your task? Have others tried to solve the same task or answer a similar research question? What are you/they doing differently?
   - *Likely challenges and mitigations*: What is hard about this task / research question? Do you have any contingency plans if the experiments do not go as planned?
2. **Related Work Section**: Include 3-4 sentence descriptions of each related paper, consisting of a total of no less than 4 papers directly relevant to the proposed project. Also mention how the problem you are working on is similar/different from these.

Notes

Outline:

1. Task / Research Question Description:
    a. I want to build a multi-varietal classifiers for Quechua texts
    b. There is currently only a classifier that does classification from Cuzco Quechua as a binary yes or no
        i. Medina paper
2. Motivation & existing work:
    a. Quechua is a low-resouce language
        i. Repurpose this intro:
            1. Quechua, an indigenous language family in the Andean region, exhibits extensive linguistic diversity with over 40 varieties spoken by approximately 8 to 10 million speakers across several countries including Argentina, Bolivia, Colombia, Ecuador, and Peru (Luykx et al. 2016; Grimes 1985; Willem 2020; Hornberger & King 2010). This diversity poses unique challenges for linguistic analysis and natural language processing (NLP) tasks, such as automatic interlinear glossing (IGT) creation and machine translation (Buys & Botha 2018; Himoro et al. 2022; Wiemerslage 2022), due to wide-ranging orthographic variations and dialectal differences (Hornberger & Limerick 2019; Limerick 2018).
    b. I want to do this task because this is a preliminary step to a larger project I am working on which is Cross-Dialectical Morphological Parsing for Quechuan Languages
        i. Here is the info from my prelim on the project:
            1. This study proposes a two-fold approach: a linguistic analysis to elucidate the typological, morphological, and orthographic differences between Quechua varieties, followed by the development of a computational model for cross-dialectal morphological parsing.
            2. The linguistic analysis involves assessing existing linguistic studies and pre-annotated corpora to uncover typological, morphological, and orthographic differences between Quechua varieties. Specifically, orthographical, morphological, and phonological analyses will be conducted to identify key variations among the varieties. These analyses will inform subsequent modules of the computational model, providing crucial insights into the linguistic characteristics of each variety.
            3. Drawing inspiration from previous works in cross-varietal processing and morphological parsing (Buys & Botha 2016; Harrat et al. 2015;

Ortega & Pillaipakkamnatt 2018; Rios & Mamani 2014; Sadhan et al. 2021; Sapp et al. 2023; Wiemerslage et al. 2022; Zalmout & Habash 2019; Zevallos et al. 2022), the computational model aims to design a neural network-based morphological parser tailored to accommodate varietal variations in Quechua morphology. Unlike previous approaches that focused on normalizing, parsing, and then denormalizing text (Harrat et al. 2015; Boujelbane et al. 2016; El Kholy & Habash 2010), this study aims to address the gap in the literature by focusing on variety identification and developing sub-modules to deal with different varieties dynamically, allowing for a more nuanced and accurate analysis of morphological structures across Quechua varieties.

4. Building upon architectures such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, or transformer models customized for morphological parsing tasks (Himoro et al. 2022), the study seeks to identify the most suitable architecture for handling Quechua varietal variations. Given the low-resource and morphologically rich nature of the task, the intention is to fine-tune pre-trained models using smaller annotated datasets from target varieties to enhance parsing accuracy and adaptability, while incorporating linguistic typological distinctions between Quechua varieties into the model's architecture and training process. By introducing these variety-specific linguistic features or constraints, as learned from the linguistic analysis section of this project, the aim is to guide the morphological parsing process and improve model generalization across varietal variations (Sandhan et al. 2021).
5. As part of the evaluation, the performance of the parser will be assessed using standardized metrics, including accuracy, precision, recall, and F1 score. The model's effectiveness in capturing morphological structures across Quechua varieties will be validated through qualitative analysis and linguistic review and will consider previous models (Buys and Botha (2016); Himoro and Lora (2022); Ortega and Pillaipakkamnatt (2018); Wiermerslage et al. 2022) as benchmarks in for evaluation.
6. Previous studies have made strides in computational processing for Quechua languages. Jiménez Medina (2013) developed an automatic text classifier for a binary differentiation between Cuzco-Quechua dialects and other Quechua dialects, while Rios and Mamani (2014) focused on morphological disambiguation and text normalization for Southern Quechua varieties. However, an unexplored aspect is direct parsing of Quechua varieties without denormalization. This study addresses this gap by focusing on variety identification and developing parsing modules tailored to different varieties. Instead of denormalizing, the proposed parser will be triggered by the specific variety being parsed, allowing for

a more nuanced and accurate analysis of morphological structures across Quechua varieties.

7. Ongoing projects aim to construct a WordNet and implement part-of-speech tagging for Quechua varieties (Vergara 2022). Additionally, initiatives like QuBERT offer valuable resources for tasks like named-entity recognition and POS tagging (Zevallos et al. 2022). These efforts in Quechuan morphological processing inform and complement this study.
8. In summary, the proposed computational modeling approach aims to conduct a thorough linguistic analysis to understand varietal differences within Quechua. By integrating existing linguistic knowledge and addressing the challenges posed by varietal variations in a low-resource setting, this approach aims to contribute to the computational processing of Quechua varieties.

3. *Likely challenges and mitigations*:
   - a. I see it being an issue that there is so little info on any of these dialects
     - i. Some resources though I could use include the article from Vegara
     - ii. Other computational resources such as the Qu-BERT
     - iii. Collection of Quechua Corpora I have found
4. Related Work
   - a. Go into the details of the Medina Paper and into other morphological parsers

From Prelim:

Main thing that this paper is about:

Jiménez Medina (2013) constructed the first automatic text classifier for Quechua dialects, utilizing attributes such as words, lemmas, bigrams, and trigrams to gain a binary distinction between Cuzco Quechua dialects and non-Cuzco Quechua dialects. This task could be built upon to include more than binary distinctions between dialects.

Some background motivation from my other paper that can be repurposed about the why of this paper:

Quechua, an indigenous language family in the Andean region, exhibits extensive linguistic diversity with over 40 varieties spoken by approximately 8 to 10 million speakers across several countries including Argentina, Bolivia, Colombia, Ecuador, and Peru (Luykx et al. 2016; Grimes 1985; Willem 2020; Hornberger & King 2010). This diversity poses unique challenges for linguistic analysis and natural language processing (NLP) tasks, such as automatic interlinear glossing (IGT) creation and machine translation (Buys & Botha 2018; Himoro et al. 2022; Wiemerslage 2022), due to wide-ranging orthographic variations and dialectal differences (Hornberger & Limerick 2019; Limerick 2018).

Corpora

Corpus Name	**Quechua Variety & Ethnologue Tag**	**Size**	**Description**
Llamacha/monolingual-quechua-iic	Southern Quechua (quy, quz, qxp, qxu)	175,408 rows	Quechua text
Sample Texts in South Conchucos Quechua, with Translations and Glosses	Southern Conchucos Quechua (qxo)	77 lines	Morphologically segmented instructional text, narrative, and interview
Llullmiwan parlana	Southern Conchucos Quechua (qxo)	2239 words	Translated Quechua-Spanish text
Conversaciones con Llullmi	Southern Conchucos Quechua (qxo)	8430 words	Translated Quechua-Spanish Text
Conversaciones con Reyna	Southern Conchucos Quechua (qxo)	9474 words	Translated Quechua-Spanish Text
Pear Story Doris Quechua	Southern Conchucos Quechua (qxo)	46 lines	Morphologically segmented, translated
Textiles Doris Flormira Quechua	Southern Conchucos Quechua (qxo)	227 lines	Morphologically segmented, translated
CC_100	Various Dialects, including Cuzco Quechua (quz)	112,852 lines	Quechua text
A speech corpus of Quechua Collao	Quechua Colloa (qxp, quz)	15 hours speech & text	Quechua speech and text for automatic dimensional emotion recognition
Cuzco Quechua Resources	Cusco Quechua (quz)	10+ databases/ docs	Various dictionaries, data, and other resources

Crubadan Corpus	Several dialects of Quechua (quv, quz, qxo, qxr, qxu, qxw, qxn, quw, qub…)	Different for each dialect	Various language corpora that will be utilized for the the quz and qxo varieties

Table 2: Collected Quechua Corpora

Citations

Text classification

- CLASIFICACIÓN POR DIALECTO DE DOCUMENTOS ESCRITOS EN QUECHUA
    - Rosemary Jiménez Medina
    - https://www.unibertsitatea.net/blogak/ixa/files/2021/04/Rosemary13-11-15Proyecto.pdf
    - Abstract
        - In this work we construct the first automatic text classifier written in Quechua. This will be a binary classifier to determine the dialect of the text (Cusco, Cusco not). They represent the characteristics of the corpus texts using as attributes: words, lemmas, bigrams, trigrams and combinations of all these. The techniques of dimensionality reduction are the stop word list and stemming. Will use different types of classification algorithms: decisions trees, rules, Support Vector Machine and Naive Bayes.
        - This work has been inspired by the text classification work performed for Basque by the IXA group at the University of the Basque Country, language like Quechua is binding.
        - The results obtained when evaluating the different classifiers are encouraging and show the advantages of pre-processes applied at different stages.

Low Resource Morphological Parsing

- Cross-Lingual Morphological Tagging for Low-Resource Languages
    - https://arxiv.org/pdf/1606.04279.pdf
    - Buys, Jan, and Jan A. Botha. "Cross-lingual morphological tagging for low-resource languages." *arXiv preprint arXiv:1606.04279* (2016).
    - Morphologically rich languages often lack the annotated linguistic resources required to develop accurate natural language pro- cessing tools. We propose models suitable for training morphological taggers with rich tagsets for

low-resource languages without using direct supervision. Our approach extends existing approaches of projecting part-of-speech tags across lan- guages, using bitext to infer constraints on the possible tags for a given word type or token. We propose a tagging model us- ing Wsabie, a discriminative embedding- based model with rank-based learning. In our evaluation on 11 languages, on av- erage this model performs on par with a baseline weakly-supervised HMM, while being more scalable. Multilingual experi- ments show that the method performs best when projecting between related language pairs. Despite the inherently lossy pro- jection, we show that the morphological tags predicted by our models improve the downstream performance of a parser by +0.6 LAS on average.

- Morphological Processing of Low-Resource Languages: Where We Are and What's Next
    - https://arxiv.org/pdf/2203.08909.pdf
    - Wiemerslage, Adam, et al. "Morphological Processing of Low-Resource Languages: Where We Are and What's Next." *arXiv preprint arXiv:2203.08909* (2022).
    - Automatic morphological processing can aid downstream natural language processing appli- cations, especially for low-resource languages, and assist language documentation efforts for endangered languages. Having long been multilingual, the field of computational mor- phology is increasingly moving towards ap- proaches suitable for languages with minimal or no annotated resources. First, we survey re- cent developments in computational morphol- ogy with a focus on low-resource languages. Second, we argue that the field is ready to tackle the logical next challenge: understand- ing a language's morphology from raw text alone. We perform an empirical study on a truly unsupervised version of the paradigm completion task and show that, while existing state-of-the-art models bridged by two newly proposed models we devise perform reason- ably, there is still much room for improvement. The stakes are high: solving this task will in- crease the language coverage of morphologi- cal resources by a number of magnitudes.

Quechua Computational Resources

- Desarrollo de recursos léxicos multi-dialécticos para el quechua
    - Nelsi Belly Melgarejo Vergara
    - Abstract:
        - Las lenguas de bajos recursos como el quechua no cuentan con recursos léxicos a pesar de ser importantes para contribuir en las investigaciones y en el desarrollo de muchas herramientas de Procesamiento de Lenguaje Natural (NLP) que se benefician o requieren de recursos de este tipo, de esa forma poder contribuir en la preservación de la lengua. El objetivo de esta inves- tigación es construir una WordNet (base de datos léxica) para las variedades quechua sureño, central, amazónico y norteño, y un un etiquetado gramatical de secuencias de palabras (POS tagging)

        para la variedad del quechua sureño. Para el desarrollo de esta investigación se recopi- ló información de los diccionarios y se creó corpus paralelo quechua - español, se implementó un algoritmo de clasificación para alinear el sentido de las palabras con el synset del signifi- cado en español para cada variedad de la lengua quechua y finalmente se creó un modelo de etiquetación gramatical basado en el modelo BERT. El score obtenido para el POS tagging de la variedad quechua sureño fue 0.85 % y para el quechua central 0.8 %.
    -

```
@article{melgarejodesarrollo,
  title={Desarrollo de recursos l{\'e}xicos multi-dial{\'e}cticos para el quechua},
  author={Melgarejo Vergara, Nelsi Belly},
  publisher={Pontificia Universidad Cat{\'o}lica del Per{\'u}}
}
```

- Introducing QuBERT: A Large Monolingual Corpus and BERT Model for Southern Quechua
    - Abstract:
        - The lack of resources for languages in the Americas has proven to be a problem for the creation of digital systems such as machine translation, search engines, chat bots, and more. The scarceness of digital resources for a lan- guage causes a higher impact on populations where the language is spoken by millions of people. We introduce the first official large combined corpus for deep learning of an indige- nous South American low-resource language spoken by millions called *Quechua*. Specifi- cally, our curated corpus is created from text gathered from the southern region of Peru where a dialect of Quechua is spoken that has not traditionally been used for digital systems as a target dialect in the past. In order to make our work repeatable by others, we also offer a public, pre-trained, BERT model called *Qu- BERT* which is the largest linguistic model ever trained for any Quechua type, not just the south- ern region dialect. We furthermore test our corpus and its corresponding BERT model on two major tasks: (1) named-entity recognition (NER) and (2) part-of-speech (POS) tagging by using state-of-the-art techniques where we achieve results comparable to other work on higher-resource languages. In this article, we describe the methodology, challenges, and re- sults from the creation of QuBERT which is on on par with other state-of-the-art multilingual models for natural language processing achiev- ing between 71 and 74% F1 score on NER and 84–87% on POS tasks.

```
@inproceedings{zevallos2022introducing,
  title={Introducing qubert: A large monolingual corpus and bert model for southern quechua},
  author={Zevallos, Rodolfo and Ortega, John and Chen, William and Castro, Richard and Bel, Nuria and Toshio, Cesar and Venturas, Renzo and Aradiel, Hilario and Melgarejo, Nelsi},
  booktitle={Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing},
  pages={1--13},
  year={2022}
}
```

Quechua Dialects

1. The interpretation of relationships among **Quechua dialects**
   a. https://www.jstor.org/stable/pdf/20006728.pdf?casa_token=VAwM1YMn7oMAAAAA:s8teIPjUQgLxvnz6Sk9yQRE74g5M7rVocpYsOzdnFyJ1U83nLA36z8GcbcyaeefcYDKIIZRYnfW00fudOCwDg7wkdtL4z05KEIxf8PEMmN80msHKkkk
   b. Grimes, Joseph E. "The interpretation of relationships among Quechua dialects." *Oceanic Linguistics Special Publication*(1985): 271-284.
   c. Helpful:
      i. Counted number of Quechua varieties
2. Communicative strategies across **Quechua** languages
   a. Luykx, Aurolyn, Fernando García Rivera, and Félix Julca Guerrero. "Communicative strategies across Quechua languages." *International Journal of the Sociology of language*2016.240 (2016): 159-191.
   b. link
   c. Helpful quotes:
      i. Quechua is the most extensive indigenous language family in the western hemi- sphere, with 6–8 million speakers across the Andean region. Over 80 % of these are in Peru and Bolivia; there are also sizable populations in Ecuador and Argentina, and some 20,000 in Colombia (Howard 2011: 192). Since the late twentieth century, the spread of mass schooling throughout South America has accelerated the shift toward Spanish; however, Quechua has endured as the dominant vernacular in many rural areas, and is also widely used in some urban centers in Bolivia and Peru.
      ii. Quechua is also among the most extensively described indigenous language families. The first published grammar of "coastal Quechua"[2] appeared in 1560; a grammar of Cuzco Quechua (by Diego González Holguín) appeared in 1607. Though it is popularly known as "the language of the Inkas", Quechua's spread across the central Andean region predates the main Inka expansion by centuries (Heggerty 2007, 2008), and several scholars assert that the Inkas' own court language was not Quechua but some variant of Aymara (Hardman 1985; Hardman de Bautista 1985; Cerrón-Palomino 2004; Heggerty 2008). In any case, bi- and trilingualism were the norm throughout the Inka empire, and language and ethnicity did not necessarily correspond (Mannheim 1991: 52). During the sixteenth and seventeenth centuries, Spanish colonizers spread a modified version of Cuzco Quechua (pertaining to the group of dialects that Torero [1974] classified as "Quechua IIc") across much of the central Andes, finding it a convenient lengua general for purposes of Christianization and colonial administration.[3] Despite regular pronouncements on the desirability of "hispanicizing" the indigenous population (especially after the indigenous

uprisings of the late eighteenth century), Quechua continued to be used in urban as well as rural areas, and even enjoyed some literary production (Itier 1995).

iii. As one would expect, given its vast geographical range, Quechua had already evolved into a highly diverse language family by the time of the Spanish invasion. Even as a medium of Inka statecraft, it was not standardized[4] nor hegemonic; rather, it formed part of a complex linguistic and cultural mosaic alongside several

iv. other languages, many of which have since disappeared (Hardman de Bautista 1985; Mannheim 1991). Extant varieties of Quechua differ widely, and are not all mutually intelligible, especially in the central Peruvian valleys. A general descrip- tion of Quechua would be practically impossible, since it encompasses a range of diversity comparable to the Romance languages. This linguistic diversity is com- monly obscured by the habitual use of the term “Quechua” (and even “the Quechua language”) to refer to all varieties indistinguishably, or by the trivializing term “dialects” (Hornberger and Colonel-Molina 2004: 10; Pearce et al. 2011a: 2). Such usage is common even among experts, although published scholarly descriptions are necessarily more precise.

v.

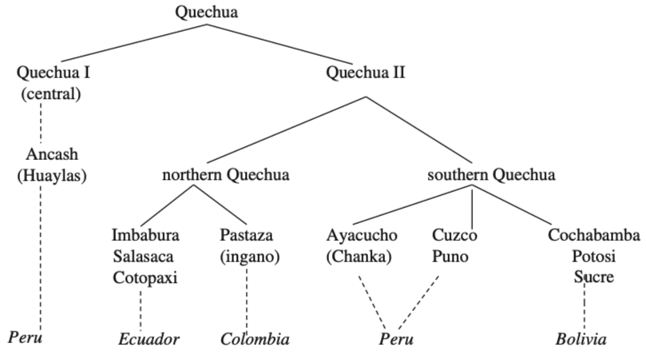

Figure 1: Quechua varieties present in initial cohort of PROEIB Andes students.

1. The variety I have information for of Quechua is from Quechua I, although a lot of the online databases focus on Quechua II, particularly for Southern Quechua

3. Morphology in Quechuan Languages
    a. https://oxfordre.com/linguistics/display/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-533
    b. Adelaar, Willem FH. "Morphology in Quechuan Languages." *Oxford Research Encyclopedia of Linguistics*. 2020.
    c. Summary
        i. Quechuan is a family of closely related indigenous languages spoken in Argentina, Bolivia, Colombia, Ecuador, and Peru, in the central part of the Andean cordilleras, in what used to be the Empire of the Incas and adjacent areas. It is divided into two main branches, commonly denominated Quechua I and II, and comprises 15 or more spoken varieties and several extinct ones that can be considered separate languages, although an exact number cannot easily be established. Quechuan shares a long and intense contact history with the neighboring Aymaran languages, but a genealogical relationship between the two families has never been demonstrated, nor a relationship with any other language family in the area. Quechuan languages are mainly agglutinative. All grammatical categories are indicated by suffixes with very few exceptions. The order in which these suffixes occur within a word form is governed by rules and combinatory restrictions that can be

rigid but not always explicable on a basis of scope and function. Portmanteau suffixes play a role in verbal inflection and in mutually interrelated domains of aspect and number in the Quechua I branch.

  ii. In Quechuan verbal derivation affixes may be semantically polyvalent, depending on the combinations in which they occur, pragmatic considerations, the nature of the root to which they are attached, their position in the affix order, and so on. Verbal derivational affixes often combine with specific verbal roots to denote meanings that are not fully predictable on the basis of the meaning of the components. Other verbal affixes never occur in such combinations. Verbal morphology and nominal morphology tend to overlap in the domain of personal reference, where subject and possessor markers are largely similar. Otherwise, the two morphological domains are almost completely separate. Not only the morphological inventories but also the formal constraints underlying the structure of verbs and nouns differ. Nominal expressions feature an elaborate but relatively instable system of case markers, some of which appear to be of recent formation. Transposition from one class to another, nominalization in particular, is indicated morphologically and occupies a central place in Quechuan grammar, particularly in interaction with case. Finally, there is a class of Independent suffixes that can be attached to members of all word classes, including adverbial elements that cannot be classified as verbs or nominals. These suffixes play a role at the organizational level of larger syntactic units, such as clauses, nominal phrases, and sentences.

4. Teachers, Textbooks, and Orthographic Choices in Quechua: Bilingual Intercultural Education in Peru and Ecuador
  a. https://brill.com/display/book/9789004298507/BP000010.xml
  b. Hornberger, Nancy H., and Nicholas Limerick. "Teachers, textbooks, and orthographic choices in Quechua: Bilingual intercultural education in Peru and Ecuador." *Perspectives on Indigenous writing and literacies*. Brill, 2019. 141-164.
  c. Helpful Quotes:
    i. One of the central paradoxes of textbook authorship in Indigenous languages is that some of those for whom the textbooks are intended find it challeng- ing to read them. Here, through examining cases of Quechua across the Andes, an example from Peru and an example from Ecuador, we consider the role of orthography in this paradox. Specifically, we show how orthographic choices of graphemes can lead to successes, as well as difficulties, in using textbooks. In addition to divides around whether one should write in Quechua more gener- ally, alphabet ideologies have complicated the creation of Quechua pedagogic materials for decades. In the case of primary grade reading materials in Quechua developed in the 1980s in the Proyecto Experimental de Educación Bilingüe (Puno Experimental Bilingual Education Project, or PEEB) in Peru, the choice to represent Quechua's vowels phonemically with three vowels or

        phonetically with five was fraught with political and pedagogical implications that changed over time. Two decades later, in Ecuador, Kukayu pedagogic readers became the signature textbook written in standardized or “Unified” Kichwa for intercultural bilin- gual schools. However, teachers and students have largely rejected them for a number of reasons, some of which include divergent orthographic standards for writing in Kichwa.

5. Authenticity and unification in Quechua language planning
    a. https://www.tandfonline.com/doi/pdf/10.1080/07908319808666564?casa_token=dlWtMf8vpH4AAAAA:NXU58YLwOd1uae__20w1Woumaz88iZscRwCjnpRay9RFt0r7Lf8SZ9b7bbv2zlcglll7hwSabljc
    b. Hornberger, Nancy H., and Kendall A. King. "Authenticity and unification in Quechua language planning." *Language Culture and Curriculum* 11.3 (1998): 390-410.
    c. Helpful quotes:
        i. With more than ten million speakers and numerous local and regional varieties, the unification and standardisation of Quechua/Quichua has been a complicated, politically charged, and lengthy process. In most Andean nations, great strides have been made towards unification of the language in recent decades. However, the process is far from complete, and multiple unresolved issues remain, at both national and local levels. A frequent sticking point in the process is the concern that the authenticity of the language will be lost in the move towards unification. This paper examines the potentially problematic tension between the goals of authenticity and unification. One case examines an orthographic debate which arose in the process of establishing an official orthography for Quechua at the national level in Peru. The second case study moves to the local level and concerns two indigenous communities in Saraguro in the southern Ecuadorian highlands where Spanish predominates but two Quichua varieties co-exist. The final section considers the implications of these debates and tensions for language planning and policy.
6. Kichwa or Quichua? Competing Alphabets, Political Histories, and Complicated Reading in Indigenous Languages
    a. https://www.journals.uchicago.edu/doi/abs/10.1086/695487?casa_token=1PCRASc2dHEAAAAA%3AGxb3zld792z9rOmbJ5q_5P3tvwOb0EmfbRAyJHLZzMzsnvYYkKLjEy7MM-uodqSQcN1CDPSneJw&journalCode=cer

b. Limerick, Nicholas. "Kichwa or Quichua? Competing alphabets, political histories, and complicated reading in Indigenous languages." *Comparative Education Review* 62.1 (2018): 103-124.

c. Over the past century, missionary educators, nation-state and academic planners, and literacy development workers have used alphabets for political ends for traditionally marginalized languages, and Native peoples have contested such planning with other alphabet proposals. Yet literacy work now often overlooks that there are multiple alphabets circulated in reading materials for the same Indigenous language. This article shows how standardization, a process long favored by academics, has been a major source of disagreement. It combines historical analysis of the politics of alphabetic literacy in Latin America with ethnographic research on Kichwa (Ecuadorian Quechua) to demonstrate how contrastive alphabets affect current literacy efforts. Distinct but overlapping alphabets create difficulties for readers in Ecuador, and alphabet histories affect how people perceive and interact with schooling materials. Sometimes just the shape of a single letter invokes emotions. Orthographies are thus bound up in histories of language contact among colonial and marginalized languages, complicating educational research, planning, and assessment's efforts to make alphabetic literacy into a monolingual, neutral, or standardized process.

Computational Morphology for Quechua

- Preliminary Results on the Evaluation of Computational Tools for the Analysis of Quechua and Aymara
    - https://aclanthology.org/2022.lrec-1.584.pdf
    - Himoro, Marcelo Yuji, and Antonio Pareja Lora. "Preliminary Results on the Evaluation of Computational Tools for the Analysis of Quechua and Aymara." *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022.
    - This paper has good overview of current tools for Quechua
    - This research has focused on evaluating the existing open-source morphological analyzers for two of the most widely spoken indigenous macrolanguages in South America, namely Quechua and Aymara. Firstly, we have evaluated their performance (precision, recall and F1 score) for the individual languages for which they were developed (Cuzco Quechua and Aymara). Secondly, in order to assess how these tools handle other individual languages of the macrolanguage, we have extracted some sample text from school textbooks and educational resources. This sample text was edited in the different countries where these macrolanguages are spoken (Colombia, Ecuador, Peru, Bolivia, Chile and Argentina for Quechua; and Bolivia, Peru and Chile for Aymara), and it includes their different standardized forms (10 individual languages of Quechua and 3 of Aymara). Processing this text by means of the tools, we have (i) calculated their coverage (number of words recognized and analyzed) and (ii) studied in detail the cases for which each tool was unable to generate any output. Finally, we discuss different ways in which these tools could be optimized, either to improve their performances or, in the specific case of Quechua, to cover more individual languages of this macrolanguage in future works as well.

- Using Morphemes from Agglutinative Languages like Quechua and Finnish to Aid in Low-Resource Translation
    - Ortega, John, and Krishnan Pillaipakkamnatt. "Using morphemes from agglutinative languages like Quechua and Finnish to aid in low-resource translation." *Proceedings of the AMTA 2018 workshop on technologies for MT of low resource languages (LoResMT 2018)*. 2018.
    - https://aclanthology.org/W18-2201.pdf
    - Quechua is a low-resource language spoken by nearly 9 million persons in South America (Hintz and Hintz, 2017). Yet, in recent times there are few published accounts of successful adaptations of machine translation systems for low-resource languages like Quechua. In some cases, machine translations from Quechua to Spanish are inadequate due to error in alignment. We attempt to improve previous alignment techniques by aligning two languages that are simi- lar due to agglutination: Quechua and Finnish. Our novel technique allows us to add rules that improve alignment for the prediction algorithm used in common machine translation systems.