Toward Creation of Ancash Quechua Lexical Resources from OCR

Johanna Cordova
INALCO ERTIM
2 rue de Lille
75007 Paris, France
johanna.cordova@inalco.fr

Damien Nouvel
INALCO ERTIM
2 rue de Lille
75007 Paris, France
damien.nouvel@inalco.fr

Abstract

The Quechua linguistic family has a limited number of NLP resources, most of them being dedicated to Southern Quechua, whereas the varieties of Central Quechua have, to the best of our knowledge, no specific resources (software, lexicon or corpus). Our work addresses this issue by producing two resources for the Ancash Quechua: a full digital version of a dictionary, and an OCR model adapted to the considered variety. In this paper, we describe the steps towards this goal: we first measure performances of existing models for the task of digitising a Quechua dictionary, then adapt a model for the Ancash variety, and finally create a reliable resource for NLP in XML-TEI format. We hope that this work will be a basis for initiating NLP projects for Central Quechua, and that it will encourage digitisation initiatives for under-resourced languages.

1 Introduction

In recent years, Quechua has become more visible in the countries where it is spoken, partly as a result of measures to strengthen its use in institutions, but also of a growing interest in these languages as a cultural element among citizens. At the same time, Quechua languages are gradually handled by NLP software. For Southern Quechua (variety of Quechua II, the most widespread linguistic family), resources already exist and many projects are experimenting large corpus digitisation to create Deep Learning models[1]. However, Quechua varieties are heterogeneous and available resources for the aforementioned variety are hardly usable for others, because of important differences in both morphology and lexicon. The present work aims at laying foundations for the development of NLP tools for another variety, the Ancash Quechua (variety of Quechua I).

The main steps described in this paper are as follows:

- We compare 3 OCR software on the task of digitising a Quechua dictionary : ABBYY Finereader, a commercial proprietary OCR; Tesseract Open Source OCR (Smith, 2007), (Smith, 2013); and GoogleDocs OCR.
- On the basis of this comparison, we use Tesseract to retrain a Quechua model to adapt it to the Ancash variety and to the specific typography of the book.
- The dictionary is fully digitised using this new model; lexical information is then gathered in a XML-TEI format.

2 State Of The Art

2.1 Ancash language and resources

Ancash is a Peruvian department located in the Central Andes, with over 30% native speakers of Quechua[2]. In this area, the Quechua varieties are relatively homogeneous and mutually intelligible, which justifies grouping them under the name Ancash Quechua. This variety is the most widely spoken of the Central Quechua linguistic branch (Q.I). However, very little data is available in digital format, and to the best our knowledge, there are none specifically prepared for NLP development.

2.1.1 Lexical resources

Since Quechua is an agglutinative language, having a lexicon would greatly facilitate the development of morphological analysis systems, which would in turn make it possible to develop useful tools for Quechuan users and the NLP community: spell checker, POS-tagger, automatic alignment of parallel corpora, etc.

Some resources are freely available in electronic format. The most widely used is probably

[1] As the OSCAR corpus https://oscar-corpus.com

[2] According to the 2017 census.

the Quechua-Spanish dictionary (Menacho López, 2005), published by the Ministry of Education, which contains 971 entries and can be queried through the online platform Qichwa 2.0[3].

An online cross-dialectal lexicon (Jacobs, 2006), featuring about 1,800 entries for Ancash, is downloadable in spreadsheet format. This format can be easily used for NLP, but the lexicon contains some redundancies, discrepancies and formatting irregularities.

The largest Ancash Quechua-Spanish dictionaries are either not officially digitised or have been published under restrictive copyright that prevent their use for NLP purposes. The main dictionaries for our variety are: Swisshelm, 1972, 399 pages, Parker et al., 1976, 311 pages; Carranza Romero, 2013, about 8,000 entries, also available as an ebook.

2.1.2 Corpora

The main corpus is in paper format only. It consists of two volumes of narratives in both Quechua and Spanish (*Cuentos y relatos en el Quechua de Huaraz*, Ramos and Ripkens, 1974), with a total of 698 pages. A digitised dictionary would be useful to automatically post-edit the OCR of this corpus (Poncelas et al., 2020).

2.2 OCR of dictionaries

The importance of digitising lexical resources for under-resourced languages has been repeatedly expressed. For the languages of the Americas, two projects are particularly similar to ours.

A off-the-shelf use of Tesseract is reported (Maxwell and Bills, 2017) to digitise 3 bilingual dictionaries (Tzeltal-English, Muinane-Spanish, Cubeo-Spanish). More specifically, authors used Tesseract's hOCR function to preserve entry's structure and infer lexical entries with associated linguistic information. A finite state transducer was used to create the lexicon from this hOCR file.

Tesseract can also be (re)trained to create dedicated models. This has been experimented for an almost extinct Canadian language (Northern Haida) (Hubert et al., 2016) for a large written corpus (100,000 words). Optimal settings discovery was conducted by training 12 models with distinct parameters. This work also experimented training the model with images generated from text using a font similar to the targeted documents, which did not prove to be efficient. The best model, trained on the original source, obtained 96.47% character rate accuracy (CRA) and a 89.03% word rate accuracy (WRA).

2.3 Quechua in OCR tools

Both Tesseract[4] and ABBYY include a pretrained Quechua model for OCR. ABBYY's model is trained on Bolivian Quechua (Q.II). The training corpus for Tesseract's model is not documented.

3 OCR of the Ancash Quechua Dictionary

3.1 Source Document

The document we digitised is a working document by the linguist Gary J. Parker, resulting from his fieldwork (Parker, 1975). It is an unpublished draft of the Ancash Quechua to Spanish dictionary mentioned above (Parker et al., 1976) This book is a list of Ancash lexemes along with their area of use (division by province), their POS, translation or gloss in Spanish, and a set of internal cross-references indicating synonyms, related terms or lectal variants. The overall structure is relatively homogeneous. The elements mentioned above are separated by blanks, but are not vertically aligned. The typography is that of the old typewriters; some typing errors remain in the document.

3.2 Typography

Ancash Quechua is written using Latin script. In the particular case of our document, the author used a phonemic spelling to represent characters whose official modern spelling is a digraph. The Table 1 shows the special characters used by Parker (in first column), their corresponding phonemes, and the graphemes commonly used today.

3.3 Methodology

3.4 Source preprocessing

After scanning the entire document in PDF format, we applied a series of pre-processing operations in order to facilitate the OCR task:

1. Cropping: cutting the file to eliminate everything that comes out of the pages. We used Gimp tool and checked for each page that the cropping did not affect the text;

[3] https://dic.qichwa.net/#/

[4] https://github.com/tesseract-ocr/tesseract

Ancash Quechua phonemes		
Character	Phoneme	Grapheme
a̱	/aː/	aa
i̱	/iː/	ii
u̱	/uː/	uu
I̱	/ʌ/	ll
c̆	/t͡ʃ/	ch
ĉ	/t͡ʂ/	tr
s̆	/ʃ/	sh
Spanish loans		
e̱, o̱	/e/, /o/ (stressed)	e, o
r̆	/ʐ/	rr

Table 1: Special characters in the dictionary

2. Conversion to greyscale and increasing contrast;

3. Conversion to high definition PNG (between 350 and 390 dpi).

The last two steps are automatically applied to the whole document thanks to a bash script, using gegl[5] and convert[6] commands.

3.5 OCR selection

In order to determine which OCR is best suited to process our document, we conducted a series of preliminary tests. We selected three of the best performing OCR software (Tafti et al., 2016), and compared their output on a set of 5 pages of the document, randomly extracted. For Tesseract's OCR, we used both Quechua and Spanish pretrained models. For ABBYY's OCR, we used the Bolivian Quechua model. GoogleDocs OCR does not allow to control any parameter. Table 2 shows the error rates for each of them.

	Tesseract	**ABBYY**	**GoogleDocs**
CER	6.64	6.43	5.26
WER	25.5	27.5	20.7

Table 2: OCR comparison on our dictionary

This evaluation shows that GoogleDocs OCR is the best performing. Many of the diacritics described in Section 3.2 are recognised, but the structure of the document is not preserved. The opposite situation occurs in the case of ABBYY. It is worth noting that the output of the latter could be greatly improved by using the numerous settings the software offers.

In addition to performances, we also took in consideration the possibility to distribute the trained model with an open licence. According to these considerations, we chose Tesseract, which gives satisfying results and allows the model to be shared.

3.5.1 Preliminary tests with Tesseract OCR

In order to have a better view of Tesseract's performance, we applied OCR on 10 PNG files, randomly extracted from the pre-processed (Section 3.4) document, using: Spanish model alone (spa FAST); Quechua model alone (que FAST); Spanish and Quechua models together (que+spa) in their compressed (FAST) and uncompressed (BEST) versions.

OCR outputs per page are concatenated into a single file, as well as corresponding gold standards. Resulting files are compared to measure Character Error Rate (CER) and Word Error Rate (WER) with the ocrevalUAtion tool[7]. Table 3 reports those evaluations.

The results show that OCR performance is relatively poor, with a word recognition accuracy (WRA) of less than 80%. The characters with diacritics presented in Section 3.2, which are absent from the character set of Quechua and Spanish models, are not recognised, making manual correction tedious. However, Tesseract offers the possibility to adapt a pre-trained model to additional fonts and characters. In the next section, we describe the training of a model specific to our book, based on Tesseract's Quechua model.

	CER	WER
spa FAST	6.23	23.20
que FAST	7.40	27.55
que+spa FAST	**5.89**	21.82
que+spa BEST	6.05	**21.29**

Table 3: Tesseract performance with pretrained models

3.6 Model training

A training corpus was built from 30 pages of the document (5,676 words, 33,687 characters).

[5] https://gegl.org/

[6] https://imagemagick.org/script/convert.php

[7] https://github.com/impactcentre/ocrevalUAtion

These pages are segmented by lines with the Tesseract hOCR tool, producing a total of 1,544 segments; each segment is then OCRised and the output is manually corrected to constitute the gold standard. The training process is done from the Quechua model. A threshold is reached at 4.379% error rate, after 4200 epochs.

3.7 Evaluation

Previous work showed that the evaluation of an OCR output depends both on the quality of the segmentation of the document and on the quality of text recognition (Karpinski et al., 2018). For this evaluation, we discarded OCR outputs whose segmentation problems affect the global structure of the page; only character recognition is thus evaluated.

Our model is evaluated on 50 randomly selected pages of the dictionary, pre-processed as described in Section 3.4. Table 4 shows CER and WER (Raw). The second score (Corr.) is computed after correction of one-off segmentation problems. The scores show an improvement of more than 3% over the Quechua+Spanish model (see Table 2 of Tesseract for the character recognition accuracy, and of 13% for the word recognition accuracy.

	Raw	Corr.
CER	**2.57**	2.42
WER	**8.19**	7.51
WER (order indep.)	6.69	5.96

Table 4: CER and WER of the Ancash Quechua model

To get a better idea of the impact of the training, we also evaluated the error rate on the characters with diacritics. Table 5 shows their volume in the training corpus (ū, ř and ĉ having only one or two occurrences, they are considered negligible) and corresponding error rates.

	Nb_{train}	Vol_{train} (%)	CER (%)
š	167	0.50	4.04
ł	157	0.47	7.83
č	148	0.44	5.11
ā	130	0.39	63.7
ī, ē, ō		<0,1	100

Table 5: Training volume and CER for special characters

Empirically, manual correction of OCR output is easier with the new model: the most frequent characters with diacritic are well recognised, and the errors are more regular, allowing in some case their automatic detection and correction. For 10 pages, we estimated an average correction time per page of 3'40.

4 Lexical Resource

During the manual correction of the OCRed text, each entry was copied into an ODS file in order to preserve the structure. The resulting file is composed of 5 columns containing the elements described in Section 3.1. Having been reviewed several times, this resource is already available online[8].

In order to distribute this resource in a format suitable for a large variety of tools, the ODS file (previously converted to CSV) is automatically converted to an XML-TEI[9] format, following the guidelines for XML encoding of dictionaries (Budin et al., 2012). The markup structure is built with the following rules :

- Spanish loans, marked in the dictionary by an asterisk before the word, are indicated by insertion of the tag `<etym>`;
- Homographs are grouped in a `<superEntry>`;
- Cross-references are marked with `<xr>`;
- Easily retrievable examples within the column corresponding to the translation or gloss are tagged with `<cit>`.

Our XML-TEI lexicon contains **3626 entries**, and is to date the largest digital resource for Ancash Quechua available for NLP and lexicometry.

5 Conclusion

The present work shows that it is relatively easy to train a new Tesseract model from an existing one, with very little data. The tests carried out on several OCRs show many that alternatives are available for this task depending on the desired output. Based on this work, we started the digitisation of a second dictionary and a corpus with the same characteristics.

[8] https://github.com/rumiwarmi/qishwar/blob/main/Diccionario%20polilectal%20-%20PARKER.ods

[9] https://tei-c.org/

Acknowledgements

We warmly thank César Itier for providing his copy of Parker's dictionary and allowing us to digitise it.

References

Gerhard Budin, Stefan Majewski, and Karlheinz Mörth. 2012. Creating lexical resources in tei p5. a schema for multi-purpose digital dictionaries. *Journal of the Text Encoding Initiative*, (3).

Francisco Carranza Romero. 2013. *Diccionario del Quechua Ancashino*. Iberoamericana Editorial Vervuert.

Isabell Hubert, Antti Arppe, Jordan Lachler, and Eddie Antonio Santos. 2016. Training & quality assessment of an optical character recognition model for Northern Haida. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3227–3234, Portorož, Slovenia. European Language Resources Association (ELRA).

Philip Jacobs. 2006. Vocabulary. http://www.runasimi.de/runaengl.htm.

Romain Karpinski, Devashish Lohani, and Abdel Belaid. 2018. Metrics for complete evaluation of ocr performance. In *IPCV'18-The 22nd Int'l Conf on Image Processing, Computer Vision, & Pattern Recognition*.

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91, Honolulu. Association for Computational Linguistics.

Leonel Alexander Menacho López. 2005. *Yachakuqkunapa Shimi Qullqa, Anqash Qichwa Shimichaw*. Ministerio de Educación, Lima, Perú.

Gary J. Parker. 1975. *Diccionario Polilectal del Quechua de Ancash*. Universidad Nacional Mayor de San Marcos.

G.J. Parker, A.C. Reyes, and A. Chávez. 1976. *Diccionario quechua, Ancash-Huailas*. Ministerio de Educación.

Alberto Poncelas, Mohammad Aboomar, Jan Buts, James Hadley, and Andy Way. 2020. A tool for facilitating ocr postediting in historical documents. *arXiv preprint arXiv:2004.11471*.

S.P. Ramos and J. Ripkens. 1974. *Cuentos y relatos en el quechua de Huaraz*. Estudios culturales benedictinos.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving ocr accuracy on early printed books by utilizing cross fold training and voting. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 423–428.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Ray Smith. 2013. History of the tesseract ocr engine: what worked and what didn't. In *Document Recognition and Retrieval XX*, volume 8658, page 865802. International Society for Optics and Photonics.

G. Swisshelm. 1972. *Un diccionario del quechua de Huaraz: quechua-castellano, castellano-quechua*. Estudios culturales benedictinos.

Ahmad P. Tafti, Ahmadreza Baghaie, Mehdi Assefi, Hamid R. Arabnia, Zeyun Yu, and Peggy Peissig. 2016. Ocr as a service: An experimental evaluation of google docs ocr, tesseract, abbyy finereader, and transym. In *Advances in Visual Computing*, pages 735–746, Cham. Springer International Publishing.