# Using Morphemes from Agglutinative Languages like Quechua and Finnish to Aid in Low-Resource Translation

**John E. Ortega** jeo10@alu.ua.es
Dept. de Llenguatges i Sistemes Informatics, Universitat d'Alacant, E-03071, Alacant, Spain

**Krishnan Pillaipakkamnatt** csckzp@hofstra.edu
Department of Computer Science, Hofstra University, Hempstead, NY 11549, USA

**Abstract**

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 similar 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.

## 1 Introduction

The NP-complete problem of translating natural languages as they are spoken by humans to machine readable text is a complex problem; yet, is partially solvable due to the accuracy of machine language translations when compared to human translations (Kleinberg and Tardos, 2005). Statistical machine translation (SMT) systems such as Moses [1], require that an algorithm be combined with enough parallel corpora, text from distinct languages that can be compared sentence by sentence, to build phrase translation tables from language models. For many European languages, the translation task of bringing words together in a sequential sentence-by-sentence format for modeling, known as word alignment, is not hard due to the abundance of parallel corpora in large data sets such as Europarl [2]. In contrast, Quechua is a language that is spoken by more than nine million people in South America (Adelaar, 2012); yet, parallel texts with Quechua in them are very scarce (Monson et al., 2006).

This paper presents an approach to address the scarcity problem of parallel corpora in Quechua. In particular, we compare our approach with a previous approach that attempted to align Quechua to German (DE) using DE as the pivot language with the final translation being Spanish (Rios et al., 2012). Generally, the consensus on solving language translation with little resources is to find more resources or use rule-based, instead of statistical-based, machine translation through employing a more controlled corpus, or set of texts like the ones presented in the Avenue project [3].

[1] http://www.statmt.org/moses/
[2] http://www.statmt.org/europarl/
[3] https://www.cs.cmu.edu/~avenue/

Additionally, scarce-resource languages like Quechua seldom have translated technical corpora like Europarl available. We attempt to employ Natural Language Processing (NLP) techniques to better align Quechua words to other, more widely studied, words in Finnish. Specifically, techniques such as pronoun identification (Lee et al., 2013), are considered by this paper to be the key strategies in attempting to find a solution to the scarcity problem.

Moses [4] builds translation tables from models and texts that are aligned from word alignment tools like Giza++ [5], the alignment module that Moses uses to pre-align text before applying heuristics to find the most likely translations from a bank of possibilities (Och and Ney, 2003). Giza++ is used to align words from sentences in parallel text. For example, the following sentence in English: "$I_1\ love_2\ you_3$" would directly align with its parallel German counterpart: "$Ich_1\ liebe_2\ dich_3$" by applying a one-to-one alignment where a position $x$ in the English sentence is directly aligned to a position $y$ in the German sentence.

The overall probability scheme used in Giza++ for the first major iteration is called the Expectation-Maximization (EM) probability (Do and Batzoglou, 2008). The focus in this paper is to review and adapt the tools that are most widely used for SMT (namely Moses and Giza++) to prove that linguistic rules that label pronouns and their counterparts can be added to obtain more accurate results for specific languages such as Quechua, a highly agglutinative language (Rios, 2011).

In the case of Quechua, most parallel texts use Spanish as the target language. Constitutional documents, plays, and poetry can be found in parallel format from various sources (Llitjós, 2007). Unfortunately, Spanish is not easily aligned to Quechua due to the complex Quechua morphology that uses suffix-based grammatical determination in order to modify words that are morphological and syntactically different from those of Spanish.

Our hypothesis is that it may be easier to take two "naturally" similar languages and compare their grammatical similarities in order to better align the languages. Most previous research attempts to translate Quechua to some other common target language, such as Spanish, have been unsuccessful due to the complexity of alignment. We leverage the abundance of Quechua–Spanish (QU–ES) corpora with the abundance of (ES–FI) text to create a final Quechua–Finnish (QU–FI) system to compare against previous work. Our aim is to modify EM algorithmic heuristics in Giza++ to achieve better Alignment Error Rates (AER) than previously published by empowering the alignment that Quechua and Finnish possess.

Moses generally uses BLEU scores to measure the preciseness of a translation. Previous work does not seem to have published Quechua translation BLEU scores because BLEU scores are normally used when there is an abundance of corpora available for the languages at hand. Our system is a hybrid rule-based and phrase-based (statistical) machine translation (MT) system for translating from Quechua to Finnish where Spanish is used as a pivot (helper) language and Giza++ is used for aligning Quechua words to Finnish words.

## 2 Related Work

Various researchers have attempted tasks like detecting entities such as nouns, verbs, and pronouns in Quechua. Several of the more important projects are based on research efforts completed in a project called the Avenue project (Llitjós, 2007). The Avenue project was created to serve as a parallel corpus project that implemented NLP tools such as a spell checker. Spell checkers, unfortunately, are not translation tools and do not attempt to map one language to another through translations. Nonetheless, spelling correctors and other editing tools can be useful for reviewing Quechua corpora's correctness of word spelling in order to ensure more precise input to a more sophisticated word alignment or machine translation tool.

[4] http://www.statmt.org/moses/

[5] http://www.statmt.org/moses/giza/GIZA++.html

Research has been completed by Rios et al. (2012) at the University of Zurich that uses the Avenue Elicitation Corpus (Llitjós, 2007). Particularly, they have performed a substantial amount of research on aligning Quechua to Spanish and vice-versa. The University of Zurich tree-banks are an attempt to annotate Quechua with the correct Part-of-Speech (POS) tags. They have taken an inflective approach by identifying suffix inflections and assigning each inflection as part of a numerical inflection group. Their work has established a good baseline research point for Quechua experimentation and is helpful with the task of translating Quechua to Spanish. However, most research completed to this date, including the Rios et al. (2012)'s research, seems to deal with the Quechua language as a whole and its translation to Spanish rather than focusing on the specific language construction and morphology. Here, we use linguistic rules to show that morphemes from Quechua to Finnish align better due to the linguistic similarity of the two languages.

Another series of morphology experiments, similar to those done at the University of Zurich, were performed by Nießen and Ney (2004). Their methodology reduced the original corpus size about ten percent resulting in only a 1.6 percent loss of translation quality while using inflectional grouping. The idea implemented by Nießen and Ney (2004) is similar to the idea researched by Rios et al. (2012) that we use for comparison in this paper. By classifying morphemes into specific inflections, or lack of inflections, groups can be formed to better statistically decide where a source word may align to a target word. The inflection idea was originally proposed by researchers at IBM (Ecker et al., 1999). Quechua is in its majority is based on inflectionally-grouped suffix morphemes. We use that phenomenon to develop a hybrid machine translation system based on Moses and Giza++. The main focus of our work is to show that rules can be applied to Quechua that will improve the error rates from Giza++ alignment results in the work performed by the University of Zurich - Parallel tree-banking Spanish-Quechua(Rios et al., 2012).

## 3 Language Specifics

Quechua is morphologically rich. Its morphology is comparable to many other European languages such as Finnish, Turkish, and even French. Quechua is a language that heavily depends on word parts, knows as morphemes, being added on as suffixes; hence, we say that Quechua is agglutinative (Rios, 2011). One example of its agglutinativity is seen with the infinitive verb in Quechua for the English verb "to grow", "wiña". The suffix "nku" is added to the word "wiña" to form the Quechua third-person plural verb, "wiña-nku", which translates to the English words "they grow". The English translation does not change the infinitive form of the word. Rather, in English, it is grammatically correct in many cases to simply add the word "they" in front of the infinitive verb to create the third-person plural form of the infinitive. It is noted that Quechua contains as many as 130 these types of suffixes (Göhring, 2014) - we deal with two of them in our work.

## 4 Methodology

We attempt to improve the Alignment Error Rates (AER) achieved by University of Zurich (Rios et al., 2012) by duplicating the results (QU–DE and QU-ES) using the same corpora and resources from their project. Then, we modify the final growth-and-reordering algorithm that Moses provides from the Giza++ alignment. It is important to note that our focus will be on the alignment ideas performed by Rios et al. (2012); therefore, we use IBM Model 1 and its lexical matching as a first step rather than focus on other, more complicated, models. All of the corpora used in this project coincide with the corpora used in the tree-banking project at the University of Zurich (Llitjós, 2007).

After duplicating the AER published by Rios et al. (2012), we create reference sentences

in Finnish. This is done by translating the previous (Spanish) reference sentences to Finnish using a Moses system trained on Europarl. Then, we manually align Quechua words to Finnish words. Slight adaptations were made to the original target reference sentences. However, the difference can be considered negligible (less than 2 words on average per sentence).

With the reference corpora created, we modify Giza++'s algorithm for alignment, the EM algorithm presented in the book by Koehn (2009), by adding practical pronoun possessive rules. After rule insertion, we rerun a new Moses (QU–FI) execution and record alignment rates by comparing the new output to our reference corpora.

### 4.1 Alignment Technique

The alignment technique we use attempts to naturally align Quechua with another language that has more readily available corpora - Finnish. Finnish has been chosen because it is quite agglutinative and, in many cases, suffix-based grammatical rules are used to modify words in the Finnish language similar to Quechua. In order to better exemplify agglutination, the example below is presented:

- Infinitive Finnish verb "to correct": korja
- Conjugate Finnish verb "to correct": korjaame (stem is korjaa)
- Infinitive Quechua verb "to correct": allinchay
- Conjugate Quechua verb "to correct": allinchaychik (stem is allinchay)

There are two main figures from the word evaluation summary table published in the parallel tree-banking paper (Rios et al., 2012) that are of most concern: 1) Spanish to Quechua words and 2) Spanish to Quechua inflectional groups. Respectively, the Alignment Error Rate (AER) achieved by the Zurich group are: 1) 85.74 and 2) 74.05. The approach taken in the parallel tree-banking paper is to use inflectional groups that will group word parts, known as lexicons (Becker, 1975), in order to translate unknown source (Spanish) words. Since Giza++ attempts reverse translations, it could be determined that a reverse translation from Quechua to Spanish would also produce around eighty percent AER. That is because the parameters used in Rios et al. (2012)'s work do not align null words and use the default methods for alignment in Giza++. The rules are not necessarily supervised because they use inflection groups(IG). An IG is a way of applying a tag to a word by annotating it according to a classification with a specific group as was done by Rios et al. (2012).

Quechua is based on a morphological structure that depends on suffixes to determine the meaning of root words that would otherwise be infinitive verbs. We modify the EM algorithm from Koehn (2009) to increase the likelihood of a word containing a desired morpheme match that has not been classified. That way matches are always done on words found in the past rather than a group of phrases. We modify the EM algorithm because other models, outside of IBM Model 1 and IBM Model2, are commonly based on fertility (Schwenk, 2007) and, thus, are not helpful when attempting to translate scarce-resource languages like Quechua. Furthermore, applying probabilities to words that cannot be aligned by a phrasal approach, where the "null" qualifier is allowed, could actually harm the output. For our purpose, which is to produce better alignment error rates than those presented in the University of Zurich parallel tree-banking project (Rios et al., 2012), all models with exception of IBM Model 1, are excluded leaving a single sentence iteration for probability purposes. While a single iteration may not be the most optimum execution operation for likelihood expectation, it serves well as a determinant for the rule-based probability. One can imagine aspects of the higher order IBM models that don't involve fertility could be useful. e.g., aspects involving distance or relative distance between matching words.

We also show that using Spanish as the pivot language for translations to Finnish makes suffixes, or morphemes, easier to align and makes inflectional grouping less necessary. Rules can be added that simply start at the end of the source word and compare them to the end of the target word. Each suffix has its own meaning and use that can be aligned using rule-based heuristics to determine the best word match. Our experiments described below show that the result of changing the target language increases the probability of lower alignment error rates.

Finnish has been chosen here for detecting pronouns through suffix identification. Pronouns in Finnish are in many cases added to the end of the stem word, or lemma, in order to signify possession or direction much like is done in Quechua. While we were unable to identify all of the suffixes with their pronouns in Quechua, we show that by adding two pronoun and possession rules we achieve higher AER.

Finnish is also ideal because rendering of Finnish sentences from Spanish sentences using a version of Moses trained on Europarl is easier than Quechua to Spanish. That makes choosing a pivot language, such as Spanish, the ideal candidate for translating the QU–ES texts to QU–FI texts and vice-versa. And, while the use of Finnish alone may be considered one of the most important factors in the alignment experiment, the focus of this paper is the adding of rules to the suffixes of both languages in order to better the AER found in previous QU–ES experiments.

Here we are working with lexical alignment between two like languages, one with low resources available. That makes a pivot language necessary. The advantage of translating by using a pivot language without a bilingual corpus available has been shown in the past by Wu and Wang (2007). By using the pivot language, we are able to translate Quechua to Finnish without having any Finnish translations directly available for Quechua. We use Finnish as the target language and Spanish as the pivot language for the alignment strategy of logical word pairing between Finnish and Quechua through their similar suffix incorporation.

## 5 Experiments

### 5.1 Tools, Corpora, and Algorithm

In order to have a clear image of how the results are achieved, we define the tools, corpora, and other necessaries of the research performed. The main tool used for attaining research results, Moses, is a combination of various tools and corpora. Apart from Moses, other auxiliary tools such as Aulex [6], an on-line translator, have been used to modify the corpora and their corresponding configuration files. Altogether, an extended amount of time was spent on preparing the input and reference sentences used for improving the alignment error rates. We use Moses for translation experiments.

There are three major phases that take place when translating a document in Moses: 1) Tokenization and Parsing, 2) Word Alignment, and 3) Phrasal and Word Tuning. For this project, the translation from Quechua to Finnish relies heavily on the first two phases above: Tokenization and Word Alignment.

Our final language model has a vocabulary from the words found in the corpora, both native and foreign, Quechua and Finnish, respectively. After preparing a model with probabilities for each word, word alignment is performed with the Giza++. We add suffix pronoun rules in order to gain higher percentages on words that are easily aligned from Quechua to Finnish.

Lastly, after word alignment is completed and saved, Moses performs final tuning and smoothing that uses statistics to determine phrase probability in the phrasal step. In our case, we only perform one alignment step of a lexicon type that compares suffixes word by word and applied commonality, or expectation, through learned words from the corpora.

As seen in Table 1, the three steps required to successfully modify rules to process

[6] http://aulex.org

Quechua to Finnish translations using Giza++ and Moses can be complex.

| **Step 1:** Tokenize and Parse | **Step 2:** Word Alignment | **Step 3:** Phrasal Tuning |
|---|---|---|
| 1. create the initial corpora<br>2. prepare corpora for word alignment<br>3. translate from Spanish to Finnish | 1. apply suffix rules<br>2. parallel word alignment from Quechua to Finnish | 1. extract word phrases<br>2. build translation table<br>3. word reordering<br>4. tuning |

Table 1: Steps for translating Quechua to Finnish in Moses using our proposed hybrid MT system

Altogether, our corpus contains 450 sentences. The SQUOIA corpora [7] from the University of Zurich tree-banking project, in its original textual format, are quite diverse and require various manual efforts in order to get quality parallel sentence translations. In order to get both the Finnish and Quechua texts in a readable format, manual reading and some command line tools are used. On-line dictionaries and publications from the following list are used create and align the parallel corpora:

- http://tatoeba.org
- http://www.runasimi.de
- http://aulex.org
- http://www.folkloredelnorte.com.ar

Apart from dictionaries, consultations from native speakers on a non-organizational basis were requested in order to review the reference sentences. But, reference sentences are not necessarily as important due to the fact that statistics, apart from the repeated occurrences of a particular word or lexicon, are not heavily used. The native speakers simply confirm that reference sentences are grammatically and logically correct.

Tools like Tixz [8] and Picaro [9] are used for alignment visualization in order to clearly view the aligned words and predict the AER (Alignment Error Rate) for the translated sentence results. In order to get results, the alignment configuration and results files have to be extracted from Moses because they are part of the overall system processing.

In order to nearly duplicate results from the University of Zurich, we execute Moses on the corpora from SQUOIA project [10] with the same parameters defined: 1) Null values are not allowed as a word 2) Fertility is not used and 3) Lexical matching is used. As an overall parameterized machine, the idea is to do word-forward matching based on the training corpora model that Giza++ creates during a single iteration. This is done by modifying the configuration file in Moses for IBM Model 1 only and adding rules directly into the IBM Model 1 EM algorithm. The basic idea of IBM Model 1 is that by applying the Chain Rule (Ambrosio and Dal Maso, 1990) of probability with two steps:

[7] https://code.google.com/archive/p/hltdi-l3/wikis/PossiblyUsefulCorpora.wiki
[8] http://texample.net/
[9] http://www.isi.edu/~riesa/software/picaro/
[10] http://a-rios.github.io/squoia/

1. Expectation application of the model and

2. Maximization estimation of the model

from the data, conversion should occur that will align native (e) words with foreign (f) words.

Since we use 446 sentences for training, probability from word references in sentence pairs alone is not enough to predict the final phrasal probability. Generally speaking, the main problem with scarce resources and statistical probability on lexical matching is the global count, or maximization of probability. The Maximization step from the EM algorithm for SMT in Moses written by Koehn (2009) takes the counts of probability and applies them at the end of execution. But, if there are few sentences in the corpus, probability cannot be skewed highly for a particular word because the amount of text that coexists in a phrasal situation is relatively low. Word alignment cannot be high (greater than fifty percent) if the sentences available are scarce. In order to maximize probability on our desired suffix rules, we modify the Em Algorithm for IBM Model 1 right before collecting counts [11].

### 5.2 Rule Addition

Modifying the Giza++ alignment algorithm for Finnish and Quechua requires a detailed understanding of Quechua and Finnish morphology. We use a few of the grammatical suffix rules from both languages that have the same meaning and convert them into rules that can be applied to the EM algorithm. Two pronoun-based rules are presented to show that the possibility for alignment error exists:

1. "chik" in Quechua to "me" in Finnish

2. "yki" in Quechua to "si" in Finnish

In order to better understand the two rules presented here that are added to the Giza++ EM algorithm, a review of both grammars and the effect of their corresponding suffixes presented above is necessary.

Rule 1 presented above is the "chik" suffix in Quechua. CHIK is a word that is a pronoun type by nature because it describes a particular part of speech: third person inclusive "we". This behavior can be seen in word like "riku-wa-n-chik". The Quechua verb "rikuy" means "to see" in English. By adding the "wa", "n", and "chik", the verb is converted into a third person group that collectively means "we see". There are exceptions to the rule. CHIK appears as "nchik" following a vowel, "ninchik" following a consonant, and "chik" elsewhere (as when it follows the "n" morpheme) (Lewis et al., 2009). Clearly, a pronoun suffix rule can be added to the EM rule in order to achieve the "we" functionality desired by adding a coefficient to the probability of the word match $p(e, f|a)$ and $p(f, e|a)$. The additional thirty-three percent of probability is added to words that fully comply with Rule 1. The inclusive third person pronoun "we" in Quechua is equivalent to the suffix in Finnish "me". The possessive suffix "mme" is compulsory in standard Finnish [12]. Finnish words that end with "me" are, thus, words that can be aligned directly with Quechua words that end with "chik". It is important to note that there are exceptions. But, considering the high error rate that currently exists (more than fifty percent), it makes sense to add this type of rule. Apart from that, the idea of explicit pronoun resolution between Quechua and Finnish has not been performed previously to our knowledge. The University of Zurich project and other projects have centralized attention on specific parts of speech and inflectional groups without specifying the specific pronoun alignment. We attempt to show

[11] line 17 of the EM algorithm on page 91 (Koehn, 2009)

[12] using standard Finnish dictionary from `http://en.wiktionary.org/wiki/-mme`

that the location of words within sentences in Quechua makes pronoun resolution somewhat possible between Quechua and Finnish. And, based on the amount of text that is available in Finnish and Spanish, pronoun resolution and specific positioning within larger corpora could possibly be attained.

Rule 2 is similar to Rule 1 in that it is based on pronoun resolution. This rule is more interesting because it directs attention to the singular second person pronoun "you". On top of that, the suffix "yki" signified that the first person is directing the root word toward the second person like the word for "I love you" in Quechua "munakuyki". The "yki" suffix is used when subject "I" does something to the direct pronoun "you", also known as the "I you" suffix (Ruiz, 2006). A direct word for word alignment from Quechua to Spanish in the example above would be almost impossible due to the amount of words in a Spanish phrase for "I love you", "Te quiero", and a Quechua word like "Munakuyki", a two-to-one comparison. Since null values are not permitted in this experiment, "munakuyki" could only be aligned to one word. Finnish does not always directly align to Quechua. For example, the Finnish equivalent for "I love you", like Spanish, is also two words, "Rakastan sinua". Nonetheless, there are more suffix-based words in Finnish that align to Quechua pronoun suffixes than in English or Spanish. That makes translating Quechua to Finnish much easier. The Finnish equivalent for the Quechua word "yki" is "si". Therefore, as is done in Rule 1, the application of probability will applied in foreign and native sentence to reflect the rule by giving a higher percentage to those words that comply with the rule. We add a 33% coefficient to rules that meet the desired requirement by modifying counts in the EM algorithm in Koehn (2009):

$$\text{count}(e|f) = \frac{t(e|f)}{total_{s(e)}} \textbf{+ .33}$$

The change will ensure that the global probability calculated for all sentences produces a higher percentage for words that are an exact lexical match for the rules proposed in here.

## 6 Results

In general, previous AERs show that when translating Quechua to Spanish, Moses and Giza++ produce high error rates as Table 2 confirms:

| **Univ. of Zurich Results** | $\boldsymbol{F_s}$ | $\boldsymbol{F_p}$ | $\boldsymbol{AER}$ |
|---|---|---|---|
| ES–QU words | 11.64 | 15.20 | 85.74 |
| Lowercase ES–QU words | 12.62 | 15.57 | 85.02 |
| ES–QU Inflectional Groups | 25.40 | 25.84 | 74.05 |
| Lowercase ES–QU Inflectional Groups | 26.53 | 26.89 | 72.95 |

Table 2: Original ES–QU results from the University of Zurich (Rios et al., 2012) where $F_s$ represents sure alignments, $F_p$ represents possible alignments, and $AER$ represents the alignment error rate.

A QU–FI execution is done with the our new, hybrid Moses system that contains two new suffix pronoun rules. As mentioned before, each counts probability has a possibility of changing by a coefficient of .33 when a rule has been met. For this experiment, there are about 12 words per sentence and one sentence per line. That means that around 6000 words have to be compared for alignment between Quechua and Finnish. Table 3 shows the hybrid rule addition results.

| **Our Hybrid Suffix-Based Results** | $F_s$ | $F_p$ | $AER$ |
|---|---|---|---|
| FI–QU words | 12.21 | 15.08 | 85.62 |
| Lowercase FI–QU words | 14.07 | 14.61 | 85.02 |
| FI–QU Inflectional Groups | 34.13 | 34.17 | 64.71 |
| Lowercase FI–QU Inflectional Groups | 34.00 | 34.13 | 61.08 |

Table 3: Hybrid suffix-based FI–QU results where $F_s$ represents sure alignments, $F_p$ represents possible alignments, and $AER$ represents the alignment error rate.

The results confirm that there are grammatical rules that can be applied directly to the suffix of a word from either language to improve alignment in the new system. That is not possible when comparing Spanish to Quechua. There are complexities when comparing the agglutinative language, Quechua, to the separated language, Spanish. There is clearly a difference between translating suffix-based translation groups in parallel word-for-word text from sentences and translating phrases that may occur in phrase-based translation with languages that are less agglutinative.

There are a large amount of suffixes that could fall under the two rules and it is clear that the AER presented may be decreased even further by classifying all of the possibilities as suffix type rules. It should be noted that, while the rules do somehow indicate supervised learning, the learning applied here is non-deterministic by nature due to the fact that grammatical construct is used as the basis for comparison for parallel words and sentences instead of a dictionary-based or single lexicon match. We leave other MT systems and forms of learning such as Zoph et al. (2016) out for this paper; but, it's would be worthwhile to try for future iterations of the system.

## 7 Conclusions

By adopting a "first-things-first" approach we overcome a number of challenges found in developing NLP Systems for resource scarce languages (Monson et al., 2006). After comparing the same reference sentences introduced in the initial experiment to our results, our work has shown successful results using suffix rules for pronouns.

The research performed has given a clear example of the possibilities of hybrid machine translation techniques with languages that have few resources available. Quechua is as an example of a low-resource spoken in various South American countries. The two rules here that are added into Giza++ are just two possibilities of the various combinations of suffixes that occur between Finnish and Quechua. Rules could be extended in Giza++ that would include all of the possibility suffixes in order to gain the best possible translation. Giza++ itself, as a word alignment tool, could be modified to accept hybrid-based rules in order to accept specific probabilities through a configuration file much like it currently does with dictionaries. The amount of possibilities that this project opens is endless. Giza++ modification is only one manner of extending this project to be applied to others.

## References

Adelaar, W. F. (2012). Modeling convergence: Towards a reconstruction of the history of quechuan–aymaran interaction. *Lingua*, 122(5):461 – 469. Language Contact and Universal Grammar in the Andes.

Ambrosio, L. and Dal Maso, G. (1990). A general chain rule for distributional derivatives. *Proceedings of the American Mathematical Society*, 108(3):691–702.

Becker, J. D. (1975). The phrasal lexicon. In *Proceedings of the 1975 Workshop on Theoretical*

*Issues in Natural Language Processing*, TINLAP '75, pages 60–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Do, C. B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897.

Ecker, D. M., Duan, L., Franz, A. M., and Horiguchi, K. (1999). Analyzing inflectional morphology in a spoken language translation system. US Patent US6442524B1.

Göhring, A. (2014). Building a spanish-german dictionary for hybrid mt. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 30–35.

Hintz, D. J. and Hintz, D. M. (2017). The evidential category of mutual knowledge in quechua. *Lingua*, 186:88–109.

Kleinberg, J. and Tardos, E. (2005). *Algorithm Design*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Lewis, M. P., Simons, G. F., Fennig, C. D., et al. (2009). *Ethnologue: Languages of the world*, volume 16. SIL international Dallas, TX.

Llitjós, A. F. (2007). *Automatic improvement of machine translation systems*. Carnegie Mellon University.

Monson, C., Llitjós, A. F., Aranovich, R., Levin, L., Brown, R., Peterson, E., Carbonell, J., and Lavie, A. (2006). Building nlp systems for two resource-scarce indigenous languages: mapudungun and quechua. *Strategies for developing machine translation for minority languages*, page 15.

Nießen, S. and Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics*, 30(2):181–204.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Rios, A. (2011). Spell checking an agglutinative language: Quechua. In *5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 51–55.

Rios, A., Volk, M., et al. (2012). Parallel treebanking spanish-quechua: how and how well do they align? *Linguistic Issues in Language Technology*, 7(1).

Ruiz, C. S. (2006). *Quechua, manual de enseñanza*, volume 4. Instituto de estudios peruanos.

Schwenk, H. (2007). Building a statistical machine translation system for french using the europarl corpus. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 189–192. Association for Computational Linguistics.

Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.