Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, January 2012

Parallel Treebanking Spanish-Quechua

How and how well do they align?

Annette Rios, Anne Göhring, Martin Volk

Published by CSLI Publications

LiLT volume 7, issue 13 January 2012

Parallel Treebanking Spanish-Quechua

How and how well do they align?

Annette Rios, Anne Göhring, Martin Volk, *Institute of Computational Linguistics, University of Zurich*

Abstract

Parallel treebanking is greatly facilitated by automatic word alignment. We work on building a trilingual treebank for German, Spanish and Quechua. We ran different alignment experiments on parallel Spanish-Quechua texts, measured the alignment quality, and compared these results to the figures we obtained aligning a comparable corpus of Spanish-German texts. This preliminary work has shown us the best word segmentation to use for the agglutinative language Quechua with respect to alignment. We also acquired a first impression about how well Quechua can be aligned to Spanish, an important prerequisite for bilingual lexicon extraction, parallel treebanking or statistical machine translation.

LiLT Volume 7, Issue 13, January 2012.
Parallel Treebanking Spanish-Quechua.

1 Introduction

Three years ago we built a first version of a parallel Spanish-Quechua treebank (Rios et al., 2009). Our current research project aims at the development of two machine translation systems. While the source language for both systems is Spanish, the target languages differ substantially: One system will translate into German, whereas the other one has the Andean indigenous language Quechua as target language. A major difficulty for this task is the limited amount of Quechua resources. The situation with parallel texts in Spanish-Quechua is even more precarious. Given these circumstances, it is worthwhile to explore alternative paths that allow the development of hybrid machine translation systems which combine the rule-based approach with statistical methods. We plan to enhance a rule-based MT system with translation rules extracted automatically from a parallel treebank.

For this reason, we build a trilingual parallel treebank with about 4000 sentences in each language. The Quechua part is currently being translated from Spanish by a professional translator in Peru.

As Quechua is a strongly agglutinative language, it is advantageous to build the syntactic trees not on complete word forms, but on smaller units. In our first version of a parallel treebank we used single morphemes as basic components of the syntactic trees annotated conforming to Role and Reference Grammar (RRG) as described in (Van Valin Jr. and Polla, 1997). This time, we intend to use dependency structures, as the annotation process with RRG is too complex and error-prone. As a further simplification, we build the dependency trees not on single morphemes, but on so called 'inflectional groups'[1], a procedure that has been described in detail for Turkish by (Eryiğit, 2007) and (Atalay et al., 2003).

Given these preconditions, we ran different alignment experiments in order to verify the usability of our morphological segmentation model for Quechua when it comes to alignment decisions. Additionally, we are interested in testing the performance of commonly used tools when applied to a typologically distant language pair for which only a small training corpus is available. For evaluation purposes, we built another parallel Spanish-German corpus[2] of the same size to compare the alignment results of both language pairs.

In this paper we present the Spanish-Quechua parallel corpus we

[1] Abbreviated in the following as *IG*s

[2] In addition to the autobiography, we selected documents from the Spanish-German treebank we are currently working on; these are reports on agriculture, education and economy.

collected so far. In section 3 we explain the morphological segmentation model for Quechua and the motivation behind the approach we chose. We describe the tools and the settings of our sentence and word alignment experiments in section 4. In the last section we present the alignment evaluation results and propose some interpretations before concluding on the insights we gained and the future prospects.

1.1 Quechua

Quechua is a group of closely related languages, spoken by 8-10 million people in Peru, Bolivia, Ecuador, Southern Colombia and the North-West of Argentina. Ethnologue[3] also lists some Quechua speakers for Chile. The Quechuan languages are divided into two main branches, Quechua I and II in terms of the Peruvian linguist (Torero, 1964). Quechua I is the more archaic group of dialects, spoken in Central Peru. It comprises a heavily fragmented dialect complex, with limited mutual comprehension between the different local varieties, although they share a number of clear common features (Adelaar and Muysken, 2004, 185). The origin of the Quechuan languages lies probably in this area (Cerrón-Palomino, 2003).

The second branch, Quechua II, comprises all the remaining Quechua dialects, spoken in Northern Peru (IIA), Ecuador and Colombia (IIB) and in Southern Peru, Bolivia and Argentina (IIC).[4] As for our project, we focus on the Quechua IIC dialect group, and within these especially on the Cuzco dialect.

Quechua is a strongly agglutinative, suffixing language. Word forms usually consist of a root and a number of suffixes, although some roots may also form a word on their own. There are more than 130 Quechua suffixes that take part in word formation, as a consequence, a Quechua root may appear in a large number of different word forms.

2 Spanish-Quechua Parallel Corpus

As the Quechua texts of our treebank are still being translated, we searched for as many alternative bilingual Spanish-Quechua texts as possible to use in our word alignment experiments. The following documents constitute the corpus of about 2500 Spanish-Quechua parallel sentences we collected so far:

- Children Rights Convention
- 3 short tales

[3] `http://www.ethnologue.com`

[4] The letters A-C stand for the linguistic distance to QI, QIIA is therefore the most akin to QI, whereas QIIC is the most divergent group respective to QI.

- Song lyrics
- Peruvian official documents:
  - Peruvian Constitution
  - National agreements (acuerdos nacionales)
- Autobiography

We found all these texts in electronic form on various internet sites, except for the last one: The autobiography of a Quechua speaking Peruvian called Gregorio Condori Mamani (Fernández and Gutiérrez, 1982). As there are only paper versions available of this text, we had to scan two books, the bilingual Spanish-Quechua edition and the German translation. We used Abbyy FineReader[5] in order to retrieve the text from the scanned images via optical character recognition (OCR). For German and Spanish, there are special language settings available, while for Quechua, this is not the case. Therefore, we selected the Spanish configuration to run the OCR procedure on the Quechua pages.

Since we want to have as little noise as possible in our data, we corrected the output of the OCR process manually before using the resulting text[6] in our alignment experiments. Surprisingly, the Quechua version contained about the same amount of errors as did the Spanish one. This is strange since the OCR system has a built-in lexicon for Spanish but none for Quechua. A possible explanation might be that the languages have similar alphabets (Quechua uses a subset of the Spanish letters) and the related orthography on one hand, and the relatively large number of common words due to the contact situation.[7]

Our corpus is heterogeneous in many respects. We might almost say that it is representative of the current use of written Quechua, except for its monolingual use in chats and blogs. As a matter of fact, the texts do not only cover various domains and genres, they also have different translation characteristics. The orally transmitted autobiography was transcribed in Quechua and translated to Spanish by the authors, this latter version being the source of the German translation. The tales were also translated from Quechua to Spanish, while the rest of the texts have been translated in the opposite direction, from Spanish to Quechua.

A major problem that always arises with Quechua texts is the lack of a written standard. Some of the texts in our corpus, namely the

[5] see `http://www.abbyy.com`

[6] We use only the first 100 Spanish sentences and their corresponding Quechua and German sentences as reference text for the alignment gold standard; see 5.1

[7] The Spanish part of the book is written in Andean Spanish, which differs considerably from standard European Spanish.

National Agreements and one of the song lyrics, are written in Ayacucho Quechua, a dialect that slightly differs from Cuzco Quechua. We 'transcribed' them into Cuzco Quechua via finite state methods. The changes include some suffix forms (e.g. progressive *-chka* in Ayacucho corresponds to *-sha* in Cuzco), but also some divergent lexical forms (e.g. *yaku* - 'water' in Ayacucho corresponds to *unu* in Cuzco). On the whole, the changes are relatively small. The Children's Rights Convention is written in the unified Southern Quechua proposed by Cerrón-Palomino (1994), which includes forms of both Ayacucho and Cuzco Quechua. This text was also transcribed to Cuzco Quechua. The rest of the texts are written in Cuzco Quechua, but use highly divergent orthographies. Additionally, some of the texts contain a considerable number of OCR errors[8]. We corrected those manually and unified the orthographies, in order to assure that the words in our small corpus had the same writing across all the texts.

Furthermore, the documents show different text structures. The tales and the autobiography are typical examples of narrative structure, consisting of loose paragraphs interspersed with dialogs. On the other hand, the tightly structured official documents like the Peruvian Constitution or the Children Rights Convention deliver good anchor points for sentence alignment, given the perfectly parallel enumeration of the individual articles. These text properties may influence the word alignment results. We believe that a thought out word segmentation is another important factor with respect to alignment.

3 Quechua Segmentation Model

Due to the rich morphological structure of the language, a single Quechua root may appear in a large number of different word forms, as noted above. The morphological complexity is a major challenge for word alignment, as even large amounts of texts will not suffice to avoid the sparse data problem.

A convenient solution is to split the word forms into groups of suffixes that can be aligned to whole words on the Spanish side. As a further benefit the segmentation of words into smaller units mitigates the sparse data problem, as individual suffix groups occur with more frequency than the corresponding complete word forms. The idea to this

[8]We did not run OCR ourselves on those texts, but judging from the kind errors they contain, it's save to assume that they were scanned and converted to text at some point. Typical OCR errors in those texts are e.g. that the apostrophe indicating glottalized stops in Quechua is written as an accent on the preceding or following letter (*llank'ay* - 'to work' written as *llankáy* or *llankày*) or the writing of 'rn' instead of 'm'. Those are definitely not errors a human writer would make.

approach has been taken from a description of the annotation process of the Turkish treebank, a language that shares many morphological features with Quechua (Eryiğit, 2007),(Atalay et al., 2003).

The segmentation model outlined in this section is purely functional and many decisions concerning the grouping of morphemes focus on providing good alignments to Spanish words. As our dependency treebank is built on the resulting morpheme groups, the annotation scheme has some influence on how words are split up, e.g. we treat case suffixes as single units, as we annotate them, in analogy to Spanish prepositions, as head of the noun they modify. The grouping of morphemes presented here is therefore highly specific to the given language pair and in no way motivated by inherent morphological structures of Quechua words.

There are 5 types of Quechua suffixes: Besides the nominalizing and verbalizing suffixes, there are many nominal and verbal derivational, respectively inflectional suffixes. Additionally, Quechua has a small set of independent suffixes (also called ambivalent suffixes in the literature). These suffixes can be attached to both verbal or nominal forms, without altering the part of speech of the given word. The position of these suffixes is at the end of the suffix sequence, their relative order is more or less fixed, though dialects show minor variations. The functions of the independent suffixes include data source, polar question marking and topic or contrast, amongst others. In combination with interrogative expressions, these suffixes may acquire special meanings (Adelaar and Muysken, 2004, p.209). In combination with demonstrative pronouns, the independent suffixes may also take the place of conjunctions, which are virtually non-existent in Quechua, unless they are borrowed from Spanish (Adelaar and Muysken, 2004).

In order to avoid the sparse data problem, and given the fact that some suffixes in Quechua correspond to Spanish words, we split the complex Quechua word forms into suffix groups, see examples 1.1 and 1.2.

(1.1)	*qati*	*-ra*	*-mu*	*-sha*	*-qti*	*-n*	*-ña*	*-s*
	to.herd	-Rptn	-Dir	-Prog	-DS	-3.Sg.Poss	-Disc	-IndE

'..when he was gone away herding (they say)..'

6 IGs:	*qatiramu*	*-sha*	*-qti*	*-n*	*-ña*	*-s*
	IG1	IG2	IG3	IG4	IG5	IG6

(1.2) *wawa* *-yki* *-kuna* *-wan*
child -2.Sg.Poss -Pl -Inst
'with your children'
3 IGs: *wawa* *-ykikuna* *-wan*
IG1 IG2 IG3

The first example, *qatiramushaqtinñas* is a subordinated clause, *-qti* is a nominalizing suffix indicating that the subject of the subordinated clause is not identical to the subject of the main clause. The subject of the subordinated clause is marked via the possessive suffix *-n*. The suffixes *-ra* and *-mu* modify the semantics of the verb root and will never be aligned to Spanish words on their own, therefore we do not separate them from the root *qati-* but instead append them to the first IG. The progressive aspect marker *-sha*, on the other hand, might correspond to a Spanish auxiliary verb used in progressive constructions, and therefore forms its own IG. In a finite verb form marked for tense and person, *-sha* would form an IG together with those suffixes. As in Spanish progressive forms the auxiliary verb bears the information of tense and person, the Quechua IG of aspect, tense and person can still be aligned to the Spanish auxiliary verb. The nominalizing suffix *-qti* can never be aligned to a Spanish word form, as it has no counterpart in this language. Nevertheless it marks a derivation boundary: As the subsequent suffixes are nominal, it cannot be part of the preceding verbal IG and consequently it forms its own IG. The same holds true for the marker of indirect evidence *-s*, it has no correspondence in Spanish, but it is definitely not part of the preceding IG either, therefore it represents its own IG.

Example 1.2, *wawaykikuna*, illustrates the separation of IGs in a purely nominal form. The plural suffix *-kuna* pluralizes the possessum (child). In a Spanish possessive phrase, plural is marked on the possessive pronoun, as well as on the noun itself (*tu-**s** hijo-**s***). Accordingly, *-kuna* (and *wawa*) can be aligned to *hijo-**s***, or it can form an IG with the possessive suffix (*-yki*), so we can align it with Spanish *sus*. We decided to use the latter approach.

4 Alignment Experiments

We used the following off-the-shelf tools in our alignment experiments:

- hunalign (1.1)
- GIZA++ (1.0.5) from the moses distribution (2010-08-13)

TABLE 1 Average number of tokens per sentence

	ES	QU		DE
		words	IGs	
ES-QU corpus	16.98	12.35	22.03	
ES-DE corpus	22.44			18.75
Gregorio corpus (gold)	19.99	12.43	24.76	20.32

- Lingua-AlignmentSet (1.1)[9]

After tokenization – the morphological segmentation in the case of Quechua – , we split the sentences automatically on paragraph boundaries and on some punctuation marks like the period, the question and interjection mark but also the colon and semicolon. Regarding the reference text used as a gold standard (see 5.1), we manually distinguished between colons after a "saying verb" introducing direct speech (these sentences were not split) and those separating long clauses. In case of short enumerations we also keep both sides of the colon together.

We chose "hunalign" instead of the "vanilla" sentence aligner as the former combines length-based with lexical information approaches. This capacity to integrate lexical information will gain importance as soon as we have enough data to extract a Spanish-Quechua bilingual dictionary.

Due to the high parallelism of our corpus documents, we aligned all the sentences at our disposal. Prior to the word alignment, we filtered the resulting aligned sentence pairs by length: if one of the aligned sentences is longer than 80 tokens, the pair is omitted. It is important to bear in mind that on the Quechua side, the token units are IGs, not word forms. Nevertheless, it is reasonable to set the same threshold for sentence length in both languages: The original Quechua texts contain less, but longer word forms. After the segmentation into inflectional groups, the average number of tokens per sentence is not significantly higher than in Spanish, see Table 1.

In order to align the words with GIZA++ we run the first 4 training steps of Moses: "prepare data, run giza, align words, lexical translation".

[9] `http://search.cpan.org/dist/Lingua-AlignmentSet`

We started from the default Moses resp. GIZA++ configuration[10] and changed the alignment symmetrization method and the maximum fertility parameters. We are aware of the limits of such a "manual tuning" approach due to the complexity of these tools. We set the following Moses resp. GIZA++ options to different configuration combinations:

- fertility ($f = 3|10$): a source word may translate, i.e. align to f target words
- alignment heuristics (al =intersect|grow-diag-final): how to combine both translation directions (1-to-n alignments) into n-to-m alignments

In order to evaluate our Quechua segmentation model, we compared the results for the alignment of Spanish to whole Quechua words with the corresponding outcome of Spanish to Quechua IGs alignment. Additionally, we conducted the experiments on lower-cased versions of our corpora.

5 Evaluation

5.1 Gold Standard

In order to prepare a gold standard for the evaluation, we used the TreeAligner annotation tool[11] to manually align the first 100 sentences of Gregorio Condori's autobiography on the word level. Whenever possible, we followed the guidelines defined for aligning syntactic trees in European Languages (Volk et al., 2009). We distinguish two alignment types according to the translation quality of the correspondence: exact alignments for words that convey the same meaning, and fuzzy alignments for words that represent approximately the same meaning,[12], e.g. consider relative clauses: While Spanish relative clauses have finite verbs marked for person and tense, the corresponding Quechua forms are nominalized verbs, and although person may be marked via possessive suffixes, this is not always the case. In this situation, we use fuzzy alignment, as the Spanish form is more specific in meaning.

For exact alignment, on the other hand, the aligned units have to convey the same meaning. They may differ in grammatical categories as, e.g. number mismatches are quite common between Spanish and Quechua. An important feature of the text used as gold standard is that in the Spanish version, the childhood memories of Gregorio are cast in past tense, whereas in the original Quechua text the story is told in the

[10] Moses: grow-diag-final alignment; GIZA++: maxfertility=10; hmmiterations=5; m1=5; m2=0; m3=3; m4=3; m5=0; m5p0=-1; m6=0

[11] see `http://kitt.cl.uzh.ch/kitt/treealigner`

[12] Other authors use the terms 'sure' and 'possible', see (Tiedemann, 2011).

FIGURE 1 Gold Standard Alignment Example

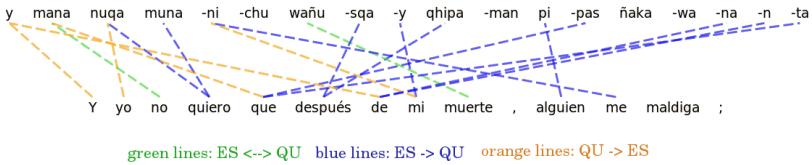

FIGURE 2 GIZA++ Alignments

unmarked Non-Future form, normally associated with present tense.[13] We decided to annotate those verb forms, despite the discrepancy in tense, with exact alignments. A verb form like Quechua *nin* -'says' and Spanish *dijo* -'said' are consequently aligned with a green line.
Figures 1 and 2 illustrate how manual alignment in the gold standard compares to GIZA++ alignments on the following sentence from Gregorio Condori's autobiography:[14]

(1.3) *..y* *mana* *nuqa* *muna* *-ni* *-chu* *wañu* *-sqa*
and Neg I want -1.Sg.Subj -Neg die -Perf
-y *qhipa* *-man* *pi* *-pas* *ñaka* *-wa* *-na*
-1.Sg.Poss behind -Dat who -Add curse -1.Obj -Purp
-n *-ta.*
-3.Sg.Poss -Acc

..y *yo* *no* *quiero* *que* *después* *de* *mi* *muerte*
and I not want that after of my death
alguien *me* *maldiga.*
someone me curse

[13] There are two tense suffixes to mark a proposition explicitly as past tense, an evidentially neutral form and a special narrative past of indirect evidentiality.
[14] For an explanation of the used abbreviations, see Appendix 1.

'..and I don't want anyone to curse me after my death.'

As can be seen in Figure 1, the nominalized verb form *ñakawananta* is split into 5 IGs, of which 3 correspond to the Spanish subjunctive verb *maldiga*, whereas the object marker *-wa* is aligned to the Spanish personal pronoun *me*. The Quechua accusative suffix, which marks the subordinated clause as complement of the main clause, is not aligned to its Spanish counterpart *que* as this correspondence is confined to constructions with complement clauses, while *que* and *-ta* do not have the same functions in other contexts. All the manual alignments in Figure 1 happen to be exact matches, represented by green lines.

Figure 2 visualizes the alignment of the same sentence pair computed by GIZA++ in our best overall configuration: Spanish words to Quechua IGs, both sides lowercased, with a maximum fertility of 3. The green lines here represent the intersection alignments, i.e. alignments present in both directions (es $\leftrightarrow$ qu); the blue lines are Spanish to Quechua alignments (es $\rightarrow$ qu); and the Quechua to Spanish alignments are orange (qu $\rightarrow$ es).

5.2 Results

The evaluation function of the Lingua-AlignmentSet tool we used can distinguish between sure and possible alignments; it thus computes 7 figures for each submitted alignment set: precision, recall and F-measure for both alignment types, and the alignment error rate (AER).[15]

The treatment of unaligned words depends on the "alignMode" option set for evaluation. With alignment mode set to "null-align", the unaligned words are forced to align with the null unit `NULL`: this affect only the results for the possible alignments, while leaving the evaluation of the secure alignments unchanged. In the "no-null-align" mode, explicit alignments to `NULL` in submitted and reference sets are suppressed for the evaluation. Our reference alignment sets (gold standard) do not contain explicit alignments to `NULL`, therefore "as-is" is equivalent to "no-null-align". Depending on the primary goal of application, a higher precision or recall may be obtained by selecting the intersection resp. "growdiagfinal" symmetrization: The former yields higher values for precision while the latter provides a better recall.[16]

Table 2 contains a summary of the results for the alignment setting that gave us the best result without "null-align".[17] If alignments to

[15] see (Och and Ney, 2003) and (Gispert et al., 2005)

[16] Conform to (Och and Ney, 2003), an even higher recall can be achieved with the union (not presented here for space reason).

[17] For the complete results, see Appendix 2.

TABLE 2 Word Alignment Evaluation Summary

Experiment max.fertility=3 alignment=growdiagfinal as-is (no null-alignment)	Fs	Fp	AER
ES-QU words	11.64	15.20	85.74
LC ES-QU words	12.62	15.57	85.02
ES-QU IGs	25.40	25.84	74.05
LC ES-QU IGs	26.53	26.89	72.95
ES-DE	39.60	40.76	59.17
LC ES-DE	40.70	41.67	58.16

LC: lowercased;
F: F-measure; AER: Alignment Error Rate
s: sure alignments; p: possible alignments

`NULL` were allowed, the Alignment Error Rate dropped below 70% for lowercased Quechua IGs, as the recall of possible alignments increases remarkably.

In general, the alignment error rate is slightly higher on the original casing than on lower-case. We observe that a lower fertility achieves better results in every setting except on Quechua words.

For similar settings, GIZA++ performs better on the Spanish-German texts than on the Spanish-Quechua corpus.

Though the Spanish-Quechua alignment is still poor, it is a satisfying result, as it confirms our hypothesis: it is "easier" to align Spanish words with Quechua IGs.

6 Conclusions

We have conducted several experiments with GIZA++ on a Spanish-Quechua, as well as a Spanish-German corpus of comparable size. As for the agglutinative language Quechua, we tested alignment not only on words, but also on inflectional groups, while for Spanish and German, no morphological information was used.

The results of the alignment experiments confirm our initial assumption that automatic alignment will lead to better results for the language pair Spanish-German. The alignment from Spanish to whole Quechua words is rather disappointing: Without alignments to `NULL`, the error rate never gets below 83%. Nevertheless, a clear improvement was achieved through the segmentation of Quechua words into inflectional groups.

We expect the performance of automatic alignment on Spanish and Quechua to improve once we have more parallel text at our disposal. So far, we have collected several books in both languages that still need to be scanned, altogether those texts contain about 20'000 parallel sentences. Furthermore, the translation of the treebank texts will provide us with 4000 additional sentence pairs. With the possibility to train GIZA++ on a larger amount of parallel texts, automatic alignment should improve.

Similar work on the language pair English-Inuktitut by Martin et al. (2003) has shown that good alignment results are possible between typologically distant languages. Inuktitut's polysynthetic morphology is much more complex than the regular agglutinative word forms of Quechua, nevertheless the authors achieved good results with a corpus of 3.4 million English words[18]. Martin et al. (2003) used the automatic alignments for dictionary expansion. They detected reliable morpheme pairs with a coverage of 72.3% of English words and a precision of 87% using pointwise mutual information (PMI). We might test their PMI method to extract word-IG pairs for our Spanish-Quechua language pair.

We are convinced that automatic alignment will improve sufficiently to facilitate the manual annotation of the parallel treebank with useful alignment suggestions.

As a further consequence, we plan to annotate the Quechua dependency trees on units of inflectional groups instead of words, as the experiments have shown a clear advantage for this approach when it comes to alignment decisions.

Acknowledgments

We would like to thank the publishers who have granted us to use part of Gregorio Condori's autobiography, as well as the many students who have contributed to its digitization. This research is funded by the Swiss National Science Foundation under grant 100015_132219/1.

[18]The Inuktitut texts contain only 1.5 million words, due to different strategies in word formation.

References

Adelaar, Willem F. H. and Pieter Muysken. 2004. *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press.

Atalay, Nart B., Kemal Oflazer, and Bilge Say. 2003. The Annotation Process in the Turkish Treebank. In *Proceedings of the 4th International Workshop on Linguistically Interpreteted Corpora (LINC)*.

Cerrón-Palomino, Rodolfo. 1994. *Quechua sureño - Diccionario Unificado*. Biblioteca Básica Peruana. Biblioteca Nacional del Perú.

Cerrón-Palomino, Rodolfo. 2003. *Lingüística Quechua*. Centro de Estudios Regionales Andinos Bartolomé de Las Casas (CBC), 2nd edn.

Eryiğit, Gülşen. 2007. ITU Treebank Annotation Tool. In *In Proceedings of the Linguistic Annotation Workshop at ACL 2007*.

Fernández, Ricardo Valderrama and Carmen Escalante Gutiérrez. 1982. *Gregorio Condori Mamani: autobiografía*. Centro Bartolomé de las Casas.

Gispert, Adrià De, Rafael Banchs, Patrik Lambert, and José B. Mariño. 2005. Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation* 39(4):267–285.

Martin, Joel, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut Parallel Corpus. In *Proceedings of the HLT-NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*.

Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1):19–51.

Rios, Annette, Anne Göhring, and Martin Volk. 2009. A Quechua-Spanish parallel treebank. In *Proceedings of the 7th Workshop on Treebanks and Linguistic Theories*. Groningen.

Tiedemann, Jörg. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Torero, Alfredo. 1964. Los dialectos quechuas. *Anales Científicos de la Universidad Agraria, Lima* (IV):446–478.

Van Valin Jr., Robert D. and Randy J. La Polla. 1997. *Syntax - Structure, Meaning and Function*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Volk, Martin, Torsten Marek, and Yvonne Samuelsson. 2009. SMULTRON (version 2.0) - The Stockholm MULtilingual parallel TReebank. http://www.cl.uzh.ch/research/paralleltreebanks_en.html. An English-German-Swedish parallel Treebank with sub-sentential alignments.

Appendix

1 Abbreviations

Acc	accusative	
Add	additive	'too, even'
Dat	dative	
Dir	directional	
Disc	discontiuative	'already'
DS	different subject	
IndE	indirect evidentiality	
Inst	instrumental	'with, by'
Neg	negation	
Obj	object	
Perf	perfect	(nominal form)
Pl	plural	
Poss	possessive suffix	
Prog	progressive	
Purp	purposive	(nominal form)
Rptn	'repentino'	'suddenly, unexpectedly'
Sg	singular	
Subj	subject	

2 Word Alignment Evaluation Results

Spanish-Quechua Results

Experiment	Ps	Rs	Fs	Pp	Rp	Fp	AER
words							
as-is							
f=10 al=growdiagfinal	8.27	28.32	12.80	11.66	29.06	16.64	84.73
f=3 al=growdiagfinal	8.29	19.52	11.64	12.03	20.61	15.20	85.74
f=10 al=intersect	9.96	3.06	4.68	14.94	3.34	5.46	94.15
f=3 al=intersect	12.13	3.70	5.67	19.25	4.27	6.99	92.67
null-align							
f=10 al=growdiagfinal	8.27	28.32	12.80	12.73	16.09	14.22	83.81
f=3 al=growdiagfinal	8.29	19.52	11.64	12.88	11.17	11.96	85.17
f=10 al=intersect	9.96	3.06	4.68	35.65	37.75	36.67	72.63
f=3 al=intersect	12.13	3.70	5.67	34.49	28.41	31.16	74.88
lower-case, as-is							
LC f=10 al=growdiagfinal	8.58	29.08	13.25	12.19	30.00	17.34	83.96
LC f=3 al=growdiagfinal	9.03	20.99	12.62	12.40	20.92	15.57	85.02
LC f=10 al=intersect	10.59	3.16	4.87	15.25	3.30	5.43	94.06
LC f=3 al=intersect	12.13	3.67	5.63	17.15	3.76	6.17	93.20
lower-case, null-align							
LC f=10 al=growdiagfinal	8.58	29.08	13.25	13.25	16.69	14.77	83.22
LC f=3 al=growdiagfinal	9.03	20.99	12.62	13.66	11.83	12.68	84.18
LC f=10 al=intersect	10.59	3.16	4.87	35.61	38.06	36.79	72.61
LC f=3 al=intersect	12.13	3.67	5.63	33.35	28.93	30.98	75.40
IGs							
as-is							
f=10 al=growdiagfinal	14.81	39.65	21.57	15.99	39.59	22.78	77.57
f=3 al=growdiagfinal	18.55	40.28	25.40	19.35	38.86	25.84	74.05
f=10 al=intersect	26.16	7.11	11.18	28.78	7.23	11.56	88.26
f=3 al=intersect	38.62	13.27	19.75	39.54	12.56	19.07	80.01
null-align							
f=10 al=growdiagfinal	14.81	39.65	21.57	17.33	22.15	19.44	76.79
f=3 al=growdiagfinal	18.55	40.28	25.40	20.03	20.46	20.24	73.70
f=10 al=intersect	26.16	7.11	11.18	34.32	47.96	40.01	72.39
f=3 al=intersect	38.62	13.27	19.75	35.05	44.09	39.05	70.76
IGs lower-case, as-is							
LC f=10 al=growdiagfinal	16.11	42.94	23.43	17.14	42.30	24.40	75.82
LC f=3 al=growdiagfinal	19.42	41.84	26.53	20.18	40.27	26.89	72.95
LC f=10 al=intersect	26.61	7.10	11.21	29.53	7.30	11.71	88.17
LC f=3 al=intersect	40.66	15.46	22.40	41.48	14.61	21.60	77.38
lower-case, null-align							
LC f=10 al=growdiagfinal	16.11	42.94	23.43	18.64	23.71	20.87	74.96
LC f=3 al=growdiagfinal	19.42	41.84	26.53	21.01	21.19	21.10	72.51
LC f=10 al=intersect	26.61	7.10	11.21	35.32	48.03	40.70	71.75
LC f=3 al=intersect	40.66	15.46	22.40	36.26	44.73	40.05	69.34

P: Precision; R: Recall; F: F-measure; AER: Alignment Error Rate
s: sure alignments; p: possible alignments

Spanish-German Results

Experiment	Ps	Rs	Fs	Pp	Rp	Fp	AER
as-is							
f=10 al=growdiagfinal	26.96	66.98	38.44	29.13	65.53	40.33	60.00
f=3 al=growdiagfinal	28.44	65.18	39.60	30.20	62.66	40.76	59.17
f=10 al=intersect	51.92	33.62	40.81	55.23	32.38	40.82	57.89
f=3 al=intersect	57.44	39.71	46.96	59.55	37.27	45.85	52.18
null-align							
f=10 al=growdiagfinal	26.96	66.98	38.44	30.73	34.84	32.65	59.11
f=3 al=growdiagfinal	28.44	65.18	39.60	31.46	32.83	32.12	58.52
f=10 al=intersect	51.92	33.62	40.81	44.11	54.49	48.75	58.65
f=3 al=intersect	57.44	39.71	46.96	46.42	56.68	51.04	55.36
lower-case, as-is							
LC f=10 al=growdiagfinal	27.59	68.78	39.38	29.55	66.69	40.95	59.22
LC f=3 al=growdiagfinal	29.23	66.98	40.70	30.88	64.05	41.67	58.16
LC f=10 al=intersect	52.52	33.10	40.61	55.24	31.52	40.14	58.34
LC f=3 al=intersect	57.69	38.94	46.49	60.10	36.72	45.59	52.53
lower-case, null-align							
LC f=10 al=growdiagfinal	27.59	68.78	39.38	30.86	34.95	32.78	58.50
LC f=3 al=growdiagfinal	29.23	66.98	40.70	32.06	33.36	32.70	57.54
LC f=10 al=intersect	52.52	33.10	40.61	44.19	55.17	49.07	58.70
LC f=3 al=intersect	57.69	38.94	46.49	47.23	57.10	51.70	54.99

P: Precision; R: Recall; F: F-measure; AER: Alignment Error Rate
s: sure alignments; p: possible alignments