A tree is a *Baum* is an *árbol* is a *sach'a*: Creating a trilingual treebank

Annette Rios, Anne Göhring

Institute for Computational Linguistics
University of Zurich
arios@ifi.uzh.ch goehring@ifi.uzh.ch

Abstract

This paper describes the process of constructing a trilingual parallel treebank. While for two of the involved languages, Spanish and German, there are already corpora with well-established annotation schemes available, this is not the case with the third language: Cuzco Quechua (ISO 639-3:quz), a low-resourced, non-standardized language for which we had to define a linguistically plausible annotation scheme first.

Keywords: Parallel treebank, Quechua, Spanish, German, dependency annotation, TrEd, PML, TIGER-XML, tree structure alignment

1. Introduction

Almost 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 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 of the corpus is currently being translated from Spanish by a professional translator in Peru.
The annotation of the Spanish-German part is finished, while for Quechua, the process is just about to start, as a suitable annotation scheme (and tool) had to be found first. 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'.

1.1. 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[1] also lists some Quechua speakers for Chile. As for our project, we focus on the dialect spoken in and around Cuzco, because this variety is relatively well described, and this circumstance considerably facilitates the development of an adequate annotation scheme. In the following, the use of the name Quechua refers explicitely to Cuzco Quechua.
Quechua is an agglutinative, suffixing language. There are more than 130 Quechua suffixes, the exact number, as well as the form of the suffixes exhibit substantial variation across dialects. There are five functional classes 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. These suffixes can be attached to both verbal or nominal forms, without altering the part of speech of the given word form. 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, 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, 208).

2. Corpus

As mentioned above, we plan to build a trilingual treebank with about 4000 sentences in each language. Parallel texts in all three languages are scarce, but not inexistent: The first text we chose for the treebank is the story of a Quechua speaking Peruvian, Gregorio Condori Mamani, whose autobiography is available in Spanish, Quechua and German, albeit in book form (Fernández and Gutiérrez, 1982). We obtained the permission from the editors to digitize part of the book for our purpose. We have integrated 500 sentences

[1] http://www.ethnologue.com

of the autobiography in our trilingual treebank.
All the remaining texts for the treebank are reports on agriculture, development aid, economy, education, media and culture. We collected these documents from the internet on the following criteria: the texts are freely available for Spanish and German, they have to be good translations of one another, and they should also contain at least 200 sentences. Additionally, we tried to limit ourselves to texts that are thematically related to Peru or at least Latin America.
We have hired a professional translator in Peru to translate these documents from Spanish to Quechua. At the time, she has translated a quarter of the corpus (about 1000 sentences).
A problem that arises when translating texts that contain vocabulary out of the 'everyday-life-domain' into Quechua is the treatment of terms or concepts that do not have a straightforward translation in this language. Almost every Quechua text contains Spanish words, specially in non-traditional Andean contexts, e.g. catholic religion, technology, economy, but also the name of animals imported to South America by the Spaniards (e.g. horse, sheep). There are two types of Spanish words in Quechua texts: Loan words and foreign words. The former are typically written according to the Quechua pronunciation and receive the same treatment as native Quechua roots, i.e. they can bear suffixes. Foreign words, on the other hand, keep the original Spanish spelling and do not bear suffixes, but instead are 'cited' with *nisqa* - 'called, said', this element then bears all the corresponding suffixes.
As for translation, we would like to have as few foreign words as possible in the Quechua texts. Nevertheless, there are cases where it would be rather confusing to 'invent' a native construction instead of using the Spanish term with *nisqa*. Every individual case has to be considered carefully, and the translation decisions should be consistent across all texts. Our translator uses a translation memory system in order to facilitate this task.
As a matter of fact, Quechua, as many other indigenous languages in South America, faces the prejudice that it serves only for the expression of rural activities and everyday life situations, but not as a medium to express scientific or technological knowledge. Therefore, we think it is important to actively produce such texts and encourage the use of Quechua in more formal contexts.
Additionally to this trilingual corpus, we have collected a smaller bilingual Spanish-Quechua corpus in order to investigate the best word segmentation for Quechua with respect to word alignment between these two languages, see (Rios et al., 2012). This corpus contains about 2500 Spanish-Quechua parallel sentences from texts of various domains and genres, e.g. the Children Rights Convention, the Peruvian Constitution, some short tales, and song lyrics, amongst others. Furthermore, we are currently digitizing another set of bilingual Spanish-Quechua books that we will add to this word aligned corpus. Once the corpus is large enough to provide reliable statistical alignments, we plan to enhance our existing parallel concordancer Align+Search[2] with the language pair Spanish - Quechua.

[2]see http://kitt.cl.uzh.ch/kitt/alignsearch/

3. Annotation

3.1. Spanish and German

The syntactic annotation process of the Spanish and German parts of the treebank is finished. For both languages, we built the syntactic trees on phrase structures. For German, we followed the well established TIGER annotation[3] scheme while for Spanish we used a simplified version of the AnCora[4] specifications developed by (Taulé et al., 2008).
Both schemes represent a hybrid form of syntactic annotation: the tree structures contain both constituent nodes and dependency labels between them. In the Spanish annotation, only the constituents directly dominated by a sentence get dependency labels: these correspond to syntactic functions like subject, object, attribute, etc. As for German, all edges are labeled, e.g. the noun of a nominal phrase is explicitly labeled as head.
We preprocess the texts before manually correcting the PoS tags and chunk phrases while annotating the sentence structures. For German we used the following tools: TreeTagger[5], Stanford Parser[6] and the TnT chunker integrated in Annotate[7]. For Spanish we use FreeLing's open-source analysis tools[8] for tagging and parsing, and Annotate to build the treebank.

3.2. Quechua

3.2.1. Annotation Scheme

The initial situation with Quechua was completely different: As there are no syntactically annotated corpora available, it was not possible to follow a previously established annotation scheme. After inspecting treebanks for other agglutinative languages, we decided to use dependency structures to represent the syntactic structure of the Quechua sentences.
Due to the agglutinative word formation in Quechua, we split the complex word forms into inflectional groups[9], an idea taken from the description of the Turkish METU-Sabanci treebank (Atalay et al., 2003; Eryiğit, 2007). These inflectional groups form the basic units of the dependency trees.
In order to develop an adequate annotation scheme from scratch we consulted several grammar books for Southern Quechua (Cusihuamán, 1976; Soto Ruiz, 1976; Cerrón-Palomino, 2003; Dedenbach-Salazar Sáenz et al., 2002). Additionally, descriptions of dependency schemes for other

[3]see http://www.ims.uni-stuttgart.de/projekte/TIGER/
[4]http://clic.ub.edu/corpus/en/ancora
[5]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
[6]see (Rafferty and Manning, 2008); Version 1.6.5 http://www-nlp.stanford.edu/software/stanford-parser-2010-11-30.tgz
[7]http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html
[8]http://nlp.lsi.upc.edu/freeling
[9]Abbreviated in the following as *IG*s

languages[10] were consulted, as considering the different approaches to dependency annotation facilitated some of the decisions concerning Quechua.
In order to give an overview on the annotation scheme for Quechua, some basic features and special cases are outlined in the following paragraphs:

VROOT As we do not include punctuation marks in our dependency trees, we introduce one non-terminal node, a virtual root (VROOT). Every punctuation mark depends directly on this virtual root as 'punc', whereas the dependency tree depends as 'sntc' (sentence) on VROOT.

Case Suffixes We consider case markers as equivalent to prepositions in languages like English (e.g. Quechua instrumental case *-wan* corresponds to English 'by, with'). In accordance with the Stanford Dependency scheme (de Marneffe and Manning, 2008) we treat case suffixes as the head of the noun they modify.

Elision of Copula The copula *kay* - 'to be' is often elided in third person contexts (see examples 3 and 4). In this case, we insert a dummy element, as the verbless clause would lack a head otherwise.

Coordination Coordinations are headless constructions by nature, therefore we arbitrarily annotate the last element as head of the preceding coordinated elements. For a head-final language like Quechua, it makes more sense to treat the last element as head instead of the first, as this is in accordance with other constructions. Coordination can be expressed through a limited set of coordinative particles[11] but also through suffixes, or both. In coordinations involving connective suffixes usually every element is morphologically marked for coordination. See Fig. 1 with the simplified annotation of the following examples:

(1) *Mana -n uywa -y -pas ni*
Not -DirE animal -1.Sg.Poss -Add nor
chakra -y -pas ka -n -chu.
field -1.Sg.Poss -Add be -3.Sg.Subj -Neg
'I don't have animals nor field.'
(lit. Neither my animal nor my field do exist')

(2) *Kapuli -ta -wan durasnu -ta -wan apa*
Capuli -Acc -Con peach -Acc -Con carry
-mu -sayki.
-Cis -1.Sg>2.Sg.Fut
'I will bring you capulis and peaches.'

(Cusihuamán, 1976, 142)

A further strategy for coordination in Quechua is the juxtaposition of two unmarked elements, e.g. *tayta mama* - 'parents' (lit. father mother) or *tuta p'unchaw* - 'night and day'.

Focus The evidential suffixes *-mi, -si* and *-cha* are usually attached to the focalized element, and thus besides their evidential function also serve as discourse markers. In yes/no-questions and negation, the interrogative/negation suffix *-chu* is attached to the focalized element (Sánchez, 2010, 47).
In their focalizing function, the evidential suffixes and *-chu* contrast with the topic markers *-qa* and *-ri*, which occupy the same slot in the suffix sequence and thus are mutually exclusive with the evidentials. Consider the following examples:

Evidential as focus marker:

(3) *Pawlucha -n wayqe -y -qa.*
Pablito -DirE/Foc brother -1.Sg.Poss -Top
'My brother is Pablito.'

(4) *Pawlucha -qa wayqe -y -mi.*
Pablito -Top brother -1.Sg.Poss -DirE/Foc
'As for Pablito, he's my brother.'

Negation suffix as focus marker:

(5) *Mana -n huwis -chu noqa -qa ka*
Not -DirE judge -Neg/Foc I -Top be
-ni.
-1.Sg.Subj
'I am not a judge (my profession is something else).'[12]

(6) *Mana -n noqa -chu huwis -qa ka*
Not -DirE I -Neg/Foc judge -Top be
-ni.
-1.Sg.Subj
'I am not the judge (the judge is someone else).'

(Cusihuamán, 1976, 93)

This morphological syncretism of two functions in a single morpheme has to be adequately represented in the dependency tree: As evidentials, they modify the clause as a whole[13], and therefore should depend on the head of the clause. Nevertheless, as focus markers, the evidentials clearly belong to the element they are attached to.
In order to represent both functions, we introduce an additional attribute 'discourse' to the terminal nodes, which gets the value 'FOCUS' if the element bears an evidential, or one of the other focus markers. The evidential itself depends on the head of the clause, see also the annotation of example 7 in Fig. 2.
The situation with the interrogative function of *-chu* is similar: As focus marker, it relates to the element it is attached

[10] English: Stanford Dependencies (de Marneffe and Manning, 2008)
Czech: Prague Dependency Treebank ((Hajičová et al., 1999; Böhmová et al., 2005)
Danish: Copenhagen Dependency Treebank (Buch-Kromann et al., 2011)

[11] e.g. *icha* - 'or' and postposition *ima* - 'also'; additionally, combinations of demonstrative pronouns with case or so-called independent suffixes may serve as clause linkers. Furthermore, Spanish borrowings like *ni* -'nor, neither' are frequently used in texts.

[12] *huwis*: from Spanish *juez* - 'judge'

[13] The occurrence of evidentials is restricted to one per clause, as there cannot be more than one data source for an utterance.

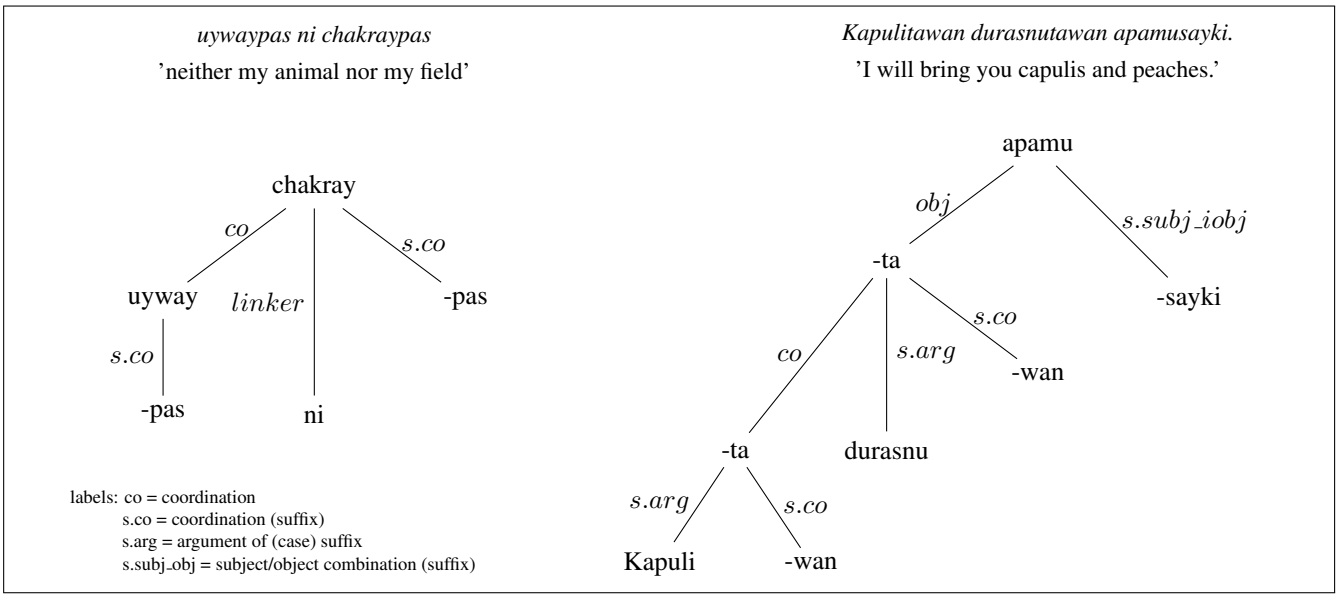

Figure 1: Coordination

to, but as interrogative suffix, it modifies the clause as a whole. As with the evidentials, we set the value for the attribute 'discourse' of the focalized element to 'FOCUS', while annotating *-chu* as direct dependent to the head of the clause.

3.2.2. Annotation process

We hired a linguistically trained Quechua speaker in Peru who will build the dependency trees on the Quechua part of the corpus. The annotation scheme elaborated so far has been tested on a limited set of sentences and may still be improved during the process. The tool of choice for the annotation is the Tree Editor (*tred*)[14] that has been used to annotate the Prague Dependency Treebank, as it is highly adaptable and very intuitive to use. Figure 2 contains the dependency tree annotated with *tred* for the following sentence from Gregorio Condori's autobiography[15]:

(7) *Aqopiya -ta -puni -n kuti -mu -y*
Acopia -Acc -Def -DirE return -Dir -Inf
-ta muna -ra -ni.
-Acc want -Pst -1.Sg.Subj
'I really want to go back to Acopia.'
(lit. 'I want my-going-back-to-Acopia')

Nodes with lexical roots contain PoS-tags (e.g. 'VRoot' - verbal root), whereas suffixes contain a tag that indicates their suffix class (e.g. 'Cas' - case suffix). Furthermore, every node has tags that contain the morphological information of the morphemes (e.g. '+Acc' - accusative case). Additionally, the Spanish translation of the lexical roots is given.
Every dependency is labeled according to the relation of the elements.

[14] http://ufal.mff.cuni.cz/~pajas/tred/

[15] Abbreviations:
Acc - accusative, Def - definitively, DirE - direct evidentiality, Dir - directional, Inf - infinitive, Pst - Past

4. Alignment

We align the correspondences in the German and Spanish parallel trees with TreeAligner, a tool developed in-house. We follow the guidelines we defined for aligning syntactic trees between English, Swedish, German and French (Volk et al., 2010), extending them to cover the Spanish specific morpho-syntactic features, e.g. clitics. All alignments represent translation correspondences that are valid independently of the given context. We distinguish two types of alignments between phrase structures, depending on the correspondence quality: exact alignment if the phrases convey the same meaning accurately, and fuzzy alignment if they provide only approximately the same information. In our example sentence, all the alignments between words and phrases are exact (green lines in Fig 3). Fuzzy alignments are set e.g. between non-finite verbs when the Spanish verb form bears a pronominal clitic, as in *comerla*, 'to eat it'.
As for Spanish and Quechua, we will convert the Spanish treebank to PML (Prague Markup Language) and use TrEd's parallel treebank extension to annotate the alignments[16]. We already have experimented on word alignments (Rios et al., 2012). We are currently writing the guidelines on how to align the different structures between Spanish and Quechua.

5. Conclusions

We have built a syntactically annotated parallel corpus of 4000 sentences in Spanish and German. In a few months, we will add a third treebank with the same texts in Cuzco Quechua. The resulting trilingual parallel treebank consists of PML (Prague Markup Language) files for all three languages, additionally the Spanish and German treebanks will be available in TIGER-XML. As for the annotation of

[16] see http://ufal.mff.cuni.cz/~pajas/tred/extensions/parallel/documentation/

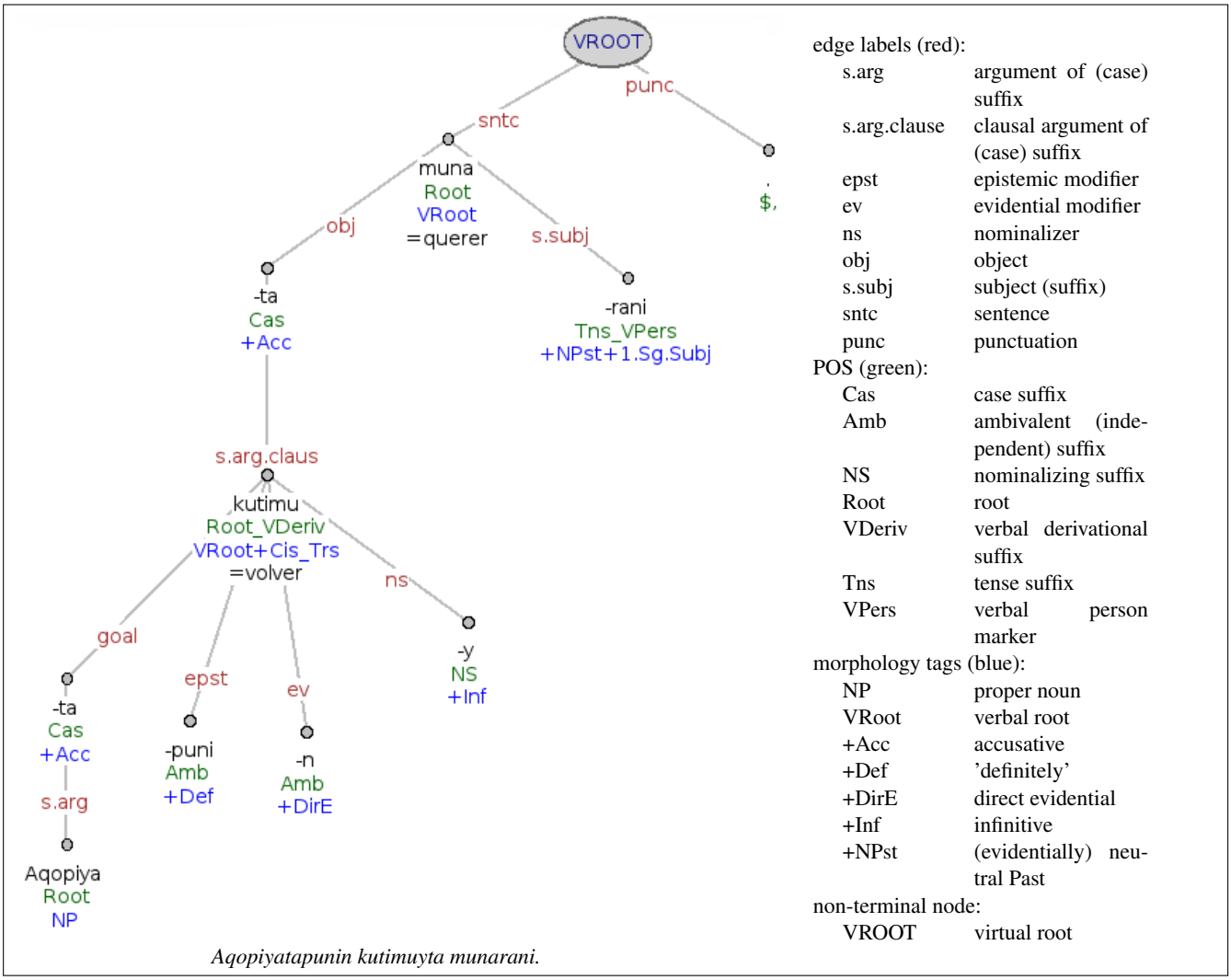

Figure 2: Quechua tree to example sentence (7)

the alignments, those are contained in separate XML files. We plan to release the Spanish-German part at mid 2012 and the Quechua part at the end of 2012 at the latest.

The main reason behind the construction of this trilingual treebank lies in its usability for machine translation: We are currently developing two rule-based MT prototypes, one in the direction Spanish to German, and another one for the translation from Spanish to (Cuzco) Quechua. Once the treebank is finished, we plan to extract translation rules and their corresponding weights from the parallel trees in order to enhance our rule-based prototype MT systems with statistical methods.

Furthermore, the published resources should also be useful for linguistic studies or language comparisons.

6. Acknowledgements

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 the annotation of the Spanish and German texts. Finally, we would like to give our special thanks to our Peruvian co-workers for the translations, the annotation as well as the linguistic consulting. This research is funded by the Swiss National Science Foundation under grant 100015_132219/1.

7. References

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

Nart B. Atalay, 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)*.

Alena Böhmová, Silvie Cinková, and Eva Hajičová. 2005. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Matthias Buch-Kromann, Morten Gylling-Jørgensen, Lotte Jelsbech Knudsen, Iørn Korzen, and Henrik Høeg Müller. 2011. The inventory of linguistic relations used in the Copenhagen Dependency Treebanks. Technical report, Copenhagen Business School, September.

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

Antonio G. Cusihuamán. 1976. *Gramática Quechua: Cuzco-Collao*. Gramáticas referenciales de la lengua

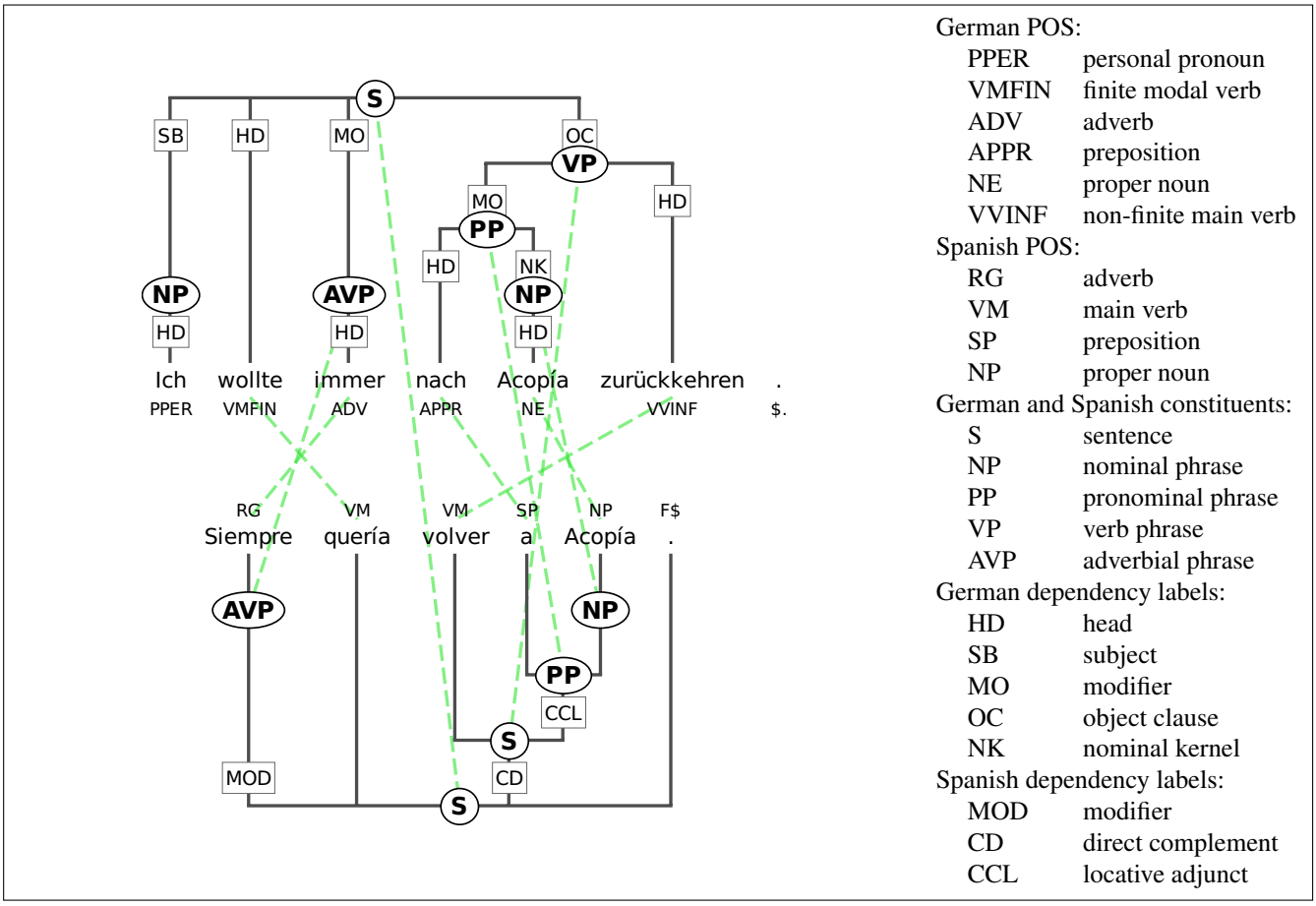

Figure 3: Aligned German and Spanish trees corresponding to Quechua example sentence (7)

quechua. Ministerio de Educación, Lima.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford dependencies manual. Technical report.

Sabine Dedenbach-Salazar Sáenz, Utta von Gleich, Roswith Hartmann, Peter Masson, Clodoaldo Soto Ruiz, and kkk. 2002. *Rimaykullayki - Unterrichtsmaterialien zum Quechua Ayacuchano*. Dietrich Reimer Verlag GmbH, Berlin, 4. edition.

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

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

Eva Hajičová, Zdeněk Kirschner, and Petr Sgall. 1999. A Manual for Analytic Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Anna Rafferty and Christopher D. Manning. 2008. Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46, Columbus, Ohio, June. Association for Computational Linguistics.

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

Annette Rios, Anne Göhring, and Martin Volk. 2012. Parallel Treebanking Spanish - Quechua: How and how well do they align? *Linguistic Issues in Language Technology (LiLT)*, 7(13), January.

Liliana Sánchez. 2010. *The Morphology and Syntax of Topic and Focus - Minimalist inquiries in the Quechua periphery*, volume 169 of *Linguistik Aktuell - Linguistics Today*. John Benjamins Publishing Company.

Clodoaldo Soto Ruiz. 1976. *Gramática Quechua: Ayacucho-Chanca*. Gramáticas referenciales de la lengua quechua. Ministerio de Educación, Lima.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

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

Martin Volk, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. 2010. SMULTRON (version 3.0) — The Stockholm MULtilingual parallel TReebank. http://www.cl.uzh.ch/research/paralleltreebanks_en.html. An English-French-German-Spanish-Swedish parallel treebank with sub-sentential alignments.