A Quechua-Spanish Parallel Treebank

Annette Rios, Anne Göhring and Martin Volk
University of Zurich
Institute of Computational Linguistics
{ozli,Anne_Goehring}@access.uzh.ch, volk@cl.uzh.ch

1 Introduction

Most treebank work in the past has focused on European and Asian languages. The Wikipedia Treebank page lists treebanks (or treebank projects) for about 20 modern European languages (ranging from Basque to Swedish), five Asian languages (Chinese, Japanese, Hindi, Korean, Thai), two ancient languages (Greek and Latin), plus Arabic and Hebrew.

Almost no treebanking work has been done on African or American indigenous languages.[1] In the past we have explored parallel treebanks for English, German and Swedish [7]. Now we would like to explore to what extent our tools and guidelines will work when we include a very different language, Quechua, for which only few NLP resources exist. Since Quechua is spoken in Latin America, Spanish as parallel language is a natural choice.

We have first compiled a parallel corpus Quechua - Spanish. We have then stepwise analyzed and annotated the Quechua and the Spanish texts. For Spanish we have used the treebanking guidelines developed by [8]. As for Quechua there were no such guidelines so that we had to experiment with finding the appropriate grammar formalism and develop our own guidelines.

In this paper we describe the characteristics of Quechua and our steps towards its morphological and syntactic annotation. We argue for Role and Reference Grammar as a suitable grammar formalism. We briefly describe how we annotated the parallel Spanish texts and demonstrate how we plan to align the Quechua with the Spanish trees.

[1] a notable exception is the work by [5]

2 Our Quechua-Spanish Corpus

Quechua is a group of closely related languages, spoken by about 8 million people in Peru, Bolivia, Ecuador, Southern Colombia and in the North of Argentina. The Quechuan Languages are divided into two subgroups, QI and QII. Quechua I is the more archaic group of dialects, spoken in Central Peru. The internal diversity between these dialects is very high, mutual intelligibility not always given. It's very likely that the origin of the Quechuan Languages lies in this area [3].

fQuechua II itself consists of three subgroups, QIIA, spoken in Northern Peru; QIIB, spoken in Ecuador and Colombia and QIIC, spoken in Southern Peru, Bolivia, and Argentina[2]. In this project, the main focus lies on the dialects of the QIIC group, and within these, especially on Cuzco and Ayacucho Quechua. The reason why QIIC was chosen for this project is very simple: For QIIC, and particularly for Cuzco and Ayacucho Quechua, there are not only by far the most linguistic descriptions at hand, but there are also more bilingual texts available than for any other dialect.

There are a lot of bilingual texts in Quechua and Spanish on the web, ranging from political texts over news to poetry and even literature. Besides these electronic texts, there are also some Quechua-Spanish printed texts and translated books, for example *Don Quijote* by Miguel de Cervantes and *Le Petit Prince* by Antoine de Saint-Exupéry. We have chosen the following texts for this project:

- the *declaration of human rights*, which is available in various Quechua dialects and contains about 100 sentences.
- some information texts and the FAQ from the website of the Peruvian *Defensoría del Pueblo*[3], which all together contain about 100 sentences.

Spanish is an official language of twenty-one countries spoken by 320 million people all around the world. Despite its geographical extension and the great regional and national diversity, it is still considered to be one language. Yet European Spanish is no longer the exclusive model for modern standard Spanish and differences can be found in pronunciation, vocabulary, and even in syntax.

The chosen text genres both influence vocabulary and sentence length. The Spanish texts of our corpus contain many juridical expressions and scarcely present any Latin American or Peruvian characteristics. The numerous adjectives, enumerations, sentence coordinations and subordinations tend to lengthen the sentences;

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

[3]The Defensoría del Pueblo is an institution that makes sure the state complies with its responsibilities for its citizens and that should also prevent the state from violating the rights of citizens.

on the other side, the short titles lower the average number of tokens per sentence to about 20.

3 Building the Quechua Treebank

3.1 Morphology

Quechua is a strongly agglutinative, suffixing language. For the morphological analyzer developed within this project we considered more than 130 suffixes. It makes sense to build Quechua syntax trees not only on whole words, but on their morphemes.[4] In order to do so, we had to develop a morphology tool that would automatically segment the words into morphemes. The challenge within this task is suffix order: Each local variety has its own preferred suffix order [3, 4], and variation in this order is not only allowed, but can sometimes even lead to a change in meaning. The solution to this problem is to group the suffixes according to their relative position in the verb. Generally, the suffixes follow this scheme:

Table 1: Suffix Order

Nominal Root	Derivation	Possession	Case		Ambivalent Suffixes
Verbal Root	Derivation	Aspect/Tense	Person	Modality	Ambivalent Suffixes

There are many ambivalent roots, that can take either verbal or nominal morphology without modification. Additionally, there are a couple of very productive nominalizing and verbalizing suffixes that can change a nominal root into a verbal one, and vice versa. Some of these suffixes can combine with each other:[5]

(1) *chinka -y.* — loss/lose -1.Sg.Poss — "My loss."

chinka -ni. — loss/lose -1.Sg.Subj — "I lose."

[4] For instance in cases without argument NPs, where subject (and object) are expressed only by verbal suffixes.

[5] Abbreviations used:

Abl	Ablative	Acc	Accusative	Ag	Nomen Agentis
AUX	Auxiliary	CLM	Clause Linkage Marker	Dat	Dative
Dem	Demonstrative Pronoun	DE	Direct Evidence	DS	Different Subject
Excl	Exclusive	Fact	Factitive	Gen	Genitive
IF	Illocutionary Force	Incl	Inclusive	Inf	Infinitive
Loc	Locative	NPst	Neutral Past	NRoot	Nominal Root
NS	Nominalizing Suffix	NUC	Nucleus	Obl	Obligation, Purpose
Perf	Perfect	Pl	Plural	PiP	Pl.Incl.Poss
Poss	Possessive Suffix	PRO	Proform	Rflx	Reflexive
Sg	Singular	Sim	Similarity	SS	Same Subject
Subj	Subject	TNS	Tense	VRoot	Verbal Root
VS	Verbalizing Suffix				

(2) *kachi* *-cha* *-sqa* *wiña* *-y* *-cha* *-ku* *-y*
salt -Fact(VS) -Perf(NS) grow -Inf(NS) -Fact(VS) -Rflx -Inf(NS)
"salted,salty" "to perpetuate oneself"

We used Xerox Finite State Tools (xfst) to build our morphological analyzer [1]. First of all, we split up Quechua Suffixes into five classes (table 2). Three out of these five classes needed further refinement, namely the N–>N, V–>V and the ambivalent suffixes.

Table 2: Suffix Classes

1	nominalizing suffixes	V –> N
2	verbalizing suffixes	N –> V
3	nominal derivational suffixes	N –> N [6]
4	verbal derivational suffixes	V –> V
5	ambivalent suffixes	N/V –> N/V

The nominal derivational suffixes (N–>N) were divided into 6, the verbal derivational suffixes (V->V) into 7 slots according to their relative position in the word. Some of these slots are iterable, i.e. more than one suffix out of a group is possible, while others are not. If more than one suffix of a given slot is present in the wordform, the relative order of these suffixes is variable, reflecting the differences between the various local varieties of the language.

The class of the ambivalent suffixes contains suffixes that are attached to nominal or verbal wordforms, without changing their part of speech. The position of these suffixes is at the end of the suffix sequence, their relative order is more or less fixed, dialects show some minor variation.

3.2 Syntax Trees

In a first attempt, we tried to build the Quechua syntax trees using phrase structures. However, Quechua poses some severe problems for this approach, and so we looked for a more appropriate grammar formalism. With respect to its complex morphological structure Quechua is similar to languages like Finnish and Estonian. Treebanks for these languages have also avoided constituent structure trees. The Estonian Arborest Treebank, for example, is based on constraint grammar which is a special type of dependency structure. [2] mention that non-finite clausal constructions pose special problems for their formalism. They solve the issue by leaving certain dependencies between subclauses underspecified. We propose that

[6] contains also possessive suffixes and case markers

RRG (Role and Reference Grammar) as described by Van Valin [9] is best suited to account for the characteristics of Quechua, including the non-finite clausal constructions, for the following reasons:

3.2.1 NP vs. VP

There is no clear-cut differentiation into NPs and VPs. Embedded clauses always contain non-finite, nominalized verbforms. These nominalized verbs are clearly nominal, they carry nominal morphology (possessive and case markers), but they also have subjects and objects, and so are clearly predicative elements. How are these forms to be treated in a constituent tree? They are no verbal phrases, but whole clauses, with their own arguments, and so they would have to be treated as clauses (S) with a nominal head. However, it seems rather unusual to have a sentence node without a finite verb in a constituent tree. A similar problem arises from the fact that the copula for 3rd person singular may be dropped, resulting in a sentence with no finite verb.
In RRG on the other hand, the predicative element PRED is not restricted to a single part of speech, in fact, any wordform can be predicative. Hence there is no problem having a CLAUSE with a noun as predicative element. The case markers of the nominal clause can be treated as Clause Linkage Markers (CLM), according to [9].

3.2.2 Headless Relative Clauses

A special form of nominalization are the so-called headless relative clauses. Such relative clauses without external head are quite common in Quechua. Consider the following example:

(3)

..ley	*-man*	*-hina*	*derecho*	*-nchik*	*-pa*	*contra*	*-n*	*-pi*	*ruwa*	*-q*
law	-Dat	-Sim	right	-1.PiP	-Gen	against	-3.Sg.Poss	-Loc	do,make	-Ag
-kuna	*-manta*	*-m*	*waqa*	*-y*	*-cha*	*-sqa*	*ka*	*-na*	*-nchik.*	
-Pl	-Abl	-DE	shelter	-Inf	-Fact	-Perf	be	-Obl	-1.PiP	

"..so that, according to the law, we are protected from those who act against our rights."[7]

[*derechonchikpa contranpi ruwaqkunamantam*] - "from [the ones] who act against our rights" is a relative clause without head. The verbal root *ruwa-* bears the nominalizing suffix *-q* (*Nomen Agentis*), followed by the plural marker *-kuna* and the case suffix *-manta*, which are clearly nominal. If there was an external

[7] Article 8 of the Declaration of Human Rights: "Everyone has the right to an effective remedy by the competent national tribunals for acts violating the fundamental rights granted him by the constitution or by law."

head, plural and case markers would be attached to the head instead.[8] So *ruwaq* is clearly a predicative element, in this case without arguments, but it could as well have. Nevertheless, its outer node cannot be a clause, since it bears a plural suffix[9], which leads to the conclusion that the whole clause has to be considered as a nominal element. The solution in RRG is to assume a NP which contains a CLAUSE with a nominal predicative element (*ruwaq*). This approach follows exactly what [9] proposes for Lakhota nominal relative clauses.[10]

3.2.3 Switch Reference

Yet another special case is Switch Reference (Clause Chaining). Consider the following sentence from the text *Llaqtaman sayapakuq -Beatriz Merino* on the website of the *Defensoría del Pueblo*.

(4)

Chay	*-ta*	*-m*	*aypa*	*-rqa*	*-ø,*	*San*	*Marcos*	*Hatun*	*Yacha*	*-y*
Dem	-Acc	-DE	achieve	-NPst	-3.Sg.Subj	San	Marcos	big	know	-Inf
Wasi	*-manta*	*"Mariano*	*Ignacio*	*Prado"*	*beca*	*-yuq*	*ka*	*-spa.*		
house	-Abl	Mariano	Ignacio	Prado	stipend	-Poss	be	-SS		

"When [she] had achieved this, [she] obtained a 'Mariano Ignacio Prado' stipend from the San Marcos University."

[*Chaytam ayparqa*] is the main clause with a finite verb, whereas [*San Marcos Yachay Wasimanta Mariano Ignacio Prado becayuq kaspa*] is the chained, nominalized clause. The problem with phrase structures is now, besides the issue whether the embedded clause has to be treated as VP or S (or even NP), the nexus type itself. To treat the embedded clause as coordinate is not accurate, yet the embedded clause shares evidentiality and tense with the main clause and it has no finite verb. But [... *becayuq kaspa*] is not subordinated either: There is no morpheme indicating the semantic relation to the main clause, nor is the embedded clause some kind of clausal object. Rather, the two clauses describe a sequence of events. In RRG, there is a third nexus type, cosubordination, that allows to represent the clauses as two clauses on their own, but sharing evidentiality (IF), see figure 1.[11]

So finally, RRG was chosen over phrase structures, although also within this framework, there is one major issue, namely the double-marking nature of Quechua. Van Valin and La Polla [9] assume that every language is either predominantly

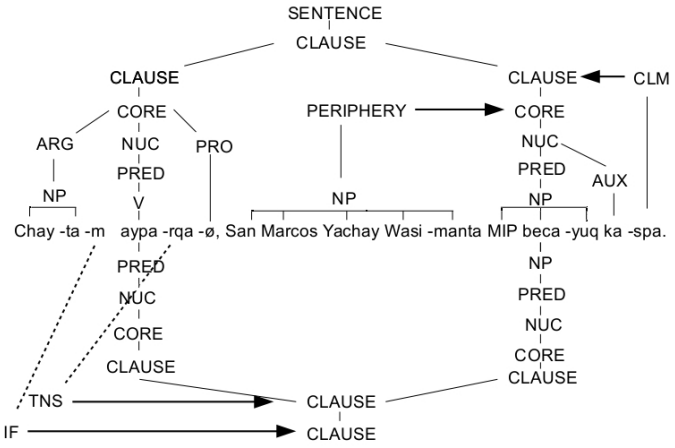
[8] e.g. with *runa*, "person": [*derechonchikpa contranpi ruwaq*] *runa**kunamanta*** - "from the persons who act against our rights"

[9] Of course the case suffix *-manta* is also a nominal suffix, but case markers can be treated as Clause Linkage Marker, see 3.2.1

[10] Van Valin's Lakhota relative clauses are internally headed, as opposed to the Quechua example, that has no head at all. But the structure is the same: an NP containing a (relative) clause.

[11] glosses see sentence 4 in 3.2.3

Figure 1: Quechua Sentence with RRG (simplified)



head- or dependent-marking. The difference is that in head-marking languages the verbal affixes are attached as PRO (Proforms) to the core, whereas the (unmarked!) NPs are considered to be outside the core and so are attached to the clause node. Considering that in Quechua the finite verb always bears person suffixes for its arguments,[12] but a sentence without argument NPs is possible, it seems more plausible to treat Quechua as a head-marking language, at least for local persons. Yet we decided to make a compromise for 3rd person objects: On account of the fact that these are never cross-referenced on the verb we chose to attach them to the CORE instead of assuming a zero-morpheme.[13] In a sentence with 1st or 2nd person object, which are cross-referenced on the verb, Quechua would be treated as head-marking, so that the core would contain the suffixes, but not the argument NPs.

We built the RRG syntax trees with the tool Annotate-3.6, which was developed to build phrase trees. RRG in fact has three levels of annotation: constituent projection, focus structure, and operator projection. This results in a three-dimensional structure. It's impossible to build such trees within Annotate-3.6, which only provides nodes, edges and secondary edges. So we decided to leave out focus structure. We then built the constituent projection without major problems, using

[12] except for 3rd person objects, which are always zero-marked

[13] As opposed to the zero-morpheme for 3rd person subject in the example sentence, which is absolutely plausible, because the 3rd person singular marker *-n* is optional after the tense suffixes *-rqa* and *-sqa*.

nodes and some labeled edges.[14] The operators were connected directly to their corresponding nodes via edges annotated with the appropriate labels. Because of the restriction in Annotate-3.6 that a word (in our case suffixes) can only be attached to one node, there were cases where secondary edges had to be used to represent operators, namely for suffixes expressing person and future tense, respectively modality all in one.

4 Building the Spanish Treebank

To syntactically annotate our Spanish corpus we used a modified version of the AnCora tagsets.[15] AnCora has three levels of annotation: a morphological, a syntactic and a semantic level. In this project, we focused on the manual syntactic annotation and kept the semantic level for future work.

On the morphological level AnCora distinguishes between the part of speech (PoS) and categories such as gender, number, case, person, time, and mode. We have simplified its morphological tagset by keeping the PoS and cutting the morphological information. Instead of having 280 different labels, we reduced the set to 33 PoS tags; then we added a label for foreign words, so that the number of PoS tags is now 34.

On the syntactic level, the AnCora corpora are annotated with constituents and functions. We reduced the constituents so that they are similar to the set of phrase constituents used in the German Negra Corpus. One of our main principles is to keep the annotation simple for the annotators. To facilitate and speed up their job they should annotate as flat as possible without losing information; in a second step we will automatically deepen the structure to obtain the same tree as if following the AnCora guidelines (similar to the deepening we have used in previous projects [6]). We thus discarded some intermediate constituent nodes, typically the nodes just under the phrases. There is another difference on the token level: AnCora has single and multiword tokens: a person's first and last name are analyzed as one token as in *(Miguel_Indurain)*. We leave the tokens separate and group them under a constituent node MPN (multi-token proper name). Other cases of multiword tokens are adverbial or conjunctive expressions like *ni_siquiera* resp. *a_pesar_de*. Again, we defined other special constituent labels to gather these complex expressions together: MTC (multi-token conjunction) and MTP (multi-token preposition). The resulting constituent tagset has 19 labels.

As for the syntactic functions, we decided to keep all the function labels in a first phase; depending on the results of this experiment, we might drop some of the

[14] PERIPHERY, PRO, AUX and ARG

[15] freely available from http://clic.ub.edu/ancora

more complex and unused labels. The function labels serve to tag only the edges under a sentence constituent S; they correspond to traditional syntactic functions (subject, object, attribute, etc.) and discourse and modality elements.

Spanish is a pro-drop language, the subject pronoun, unless emphatically used, is normally omitted. In this case, the sentence structure simply lacks a subject function. When the subject of a coordinated or subordinated sentence is elliptical, a secondary edge connects the existing subject to that sentence's constituent node.

To solve the problem of multiword tokens, as we did with the multi-token constituent nodes MPN, MTC and MTP, we defined a function SVC (support verb construction) to label the edges of the elements belonging to a light verb expression like *tener en cuenta*.

5 Aligning Quechua to Spanish

We used the Stockholm TreeAligner[16] for the alignment between the trees. Aligning Quechua to Spanish is a difficult task since the syntactic structures of the two languages differ a lot:

- Spanish uses prepositions, whereas Quechua almost exclusively uses suffixes.
- Different grammatical properties are encoded: for example, Spanish marks definiteness of NPs via articles, whereas Quechua doesn't mark definiteness, but instead marks a NP as being the topic or focus of the clause.
- Quechua uses evidential suffixes to mark the source of knowledge for each proposition, Spanish lacks a comparable category.

Often, the texts are not translated literally; the meaning is given, but with different structures. Even worse, corresponding information is often split up between various sentences. For these reasons, it is difficult to find exact alignments. Often, only fuzzy alignments were possible, if any alignment at all. Figure 2 shows an example of a Spanish sentence aligned to a Quechua subordinate clause, red lines meaning fuzzy, green lines meaning exact alignments[17] (translation see below).

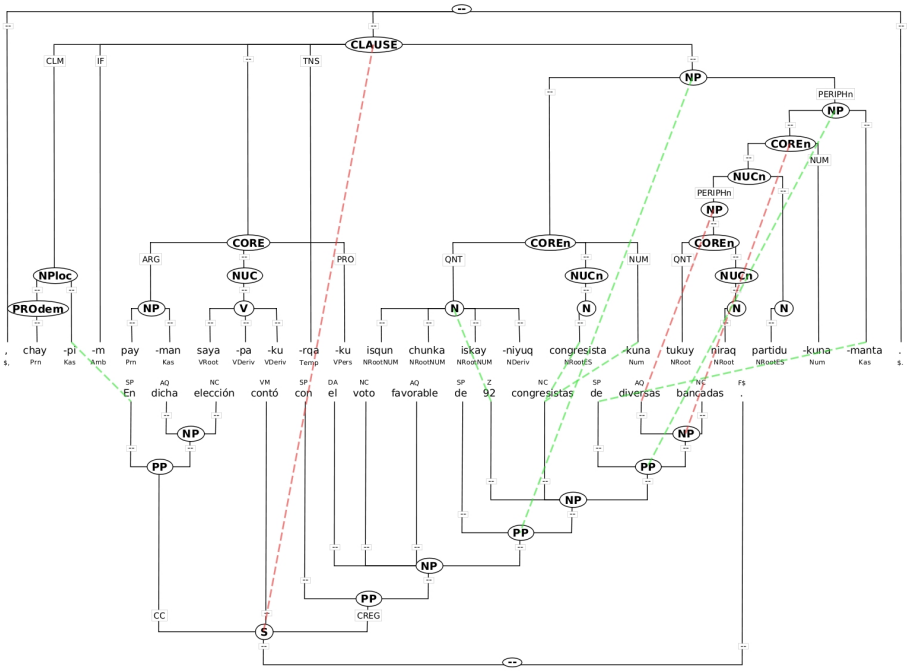
The two sentences differ in the way they express the same proposition. The literal translations would be:

- Spanish: "In the mentioned election, [she] counted with the favourable votes of 92 parliamentarians from diverse factions."

[16]The TreeAligner is available free of charge from: http://dev.ling.su.se/treealigner

[17]The Quechua main clause was cut out in this figure, for lack of space.

Figure 2: Alignment



- Quechua: “In this, [for] her stood up 92 parliamentarians from all sorts (colors) of parties.”

As you can see in Figure 2, we chose to align suffixes with prepositions when they convey the same meaning and would be good translations in other contexts too, as for example Spanish *en* and Quechua *-pi* (locative). On the other hand, *en dicha elección* - “In the mentioned election” and the corresponding Quechua part *chaypim* - “In this” could not be aligned because *en dicha elección* wouldn’t be a translation for *chaypim* in other contexts. Additionally, since the Quechua clause lacks the information conveyed by the PP *en dicha elección*, the sentence-to-clause alignment is only fuzzy (red lines). Contrary to this, the Spanish PP *de diversas bancadas* and the Quechua NP *tukuy niraq partidukunamanta* were aligned as exact matches: the internal structure is different, but the meaning conveyed is the same.

As a result of splitting up the Quechua words to their roots and suffixes, there are many multiple alignments from one Spanish word to more than one Quechua

token. For instance the Spanish word *congresistas* corresponds exactly to the Quechua *congresista* and *-kuna*[18]. In such cases, we allowed for exact multiple alignments (green lines).

6 Conclusions

We have found more bilingual texts Quechua-Spanish than we had expected. Since Quechua is a strongly agglutinative language we have decided to annotate the Quechua treebank on morphemes rather than words. This allows us to link morpho-syntactic information precisely to its source. In order to split the Quechua words into morphemes we have built a morphological analyzer based on standard finite state technology.

We realized that building phrase structure trees over Quechua sentences does not capture the characteristics of the language. We have therefore chosen Role and Reference Grammar. By using nodes, edges and secondary edges in our annotation tool we were able to represent the most important aspects of Role and Reference syntax for Quechua sentences. In order to represent all three dimensions of this formalism we will need to adapt our annotation and alignment tools.

So far, we have built the syntax structures for Quechua completely manually (after the automatic morpheme splitting). In the future we will integrate Part-of-Speech tagging and shallow parsing into the process. We will also work with alignment suggestions once we have reached a sufficiently large parallel treebank for training.

References

[1] Kenneth R. Beesley and Lauri Karttunen. *Finite State Morphology*. CSLI Publications, 2003.

[2] Eckhard Bick, Heli Uibo, and Kaili Müürisep. Arborest - a Growing Treebank of Estonian. In Henrik Holmboe, editor, *Nordisk Sprogteknologi. Nordic Language Technology. Å rbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Museum Tusculanums Forlag, Copenhagen, 2004.

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

[4] Antonio G. Cusihuamán. *Gramática Quechua: Cuzco-Collao*. Gramáticas referenciales de la lengua quechua. Ministerio de Educación, 1976.

[18] *kuna* is the suffix indicating plural, just as Spanish *-s*.

[5] C. Monson, Ariadna Font Llitjos, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. Building nlp systems for two resource-scarce indigenous languages: Mapudungun and quechua. In *Proc. of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006.

[6] Yvonne Samuelsson and Martin Volk. Automatic node insertion for treebank deepening. In *Proc. of 3rd Workshop on Treebanks and Linguistic Theories*, Tübingen, December 2004.

[7] Yvonne Samuelsson and Martin Volk. Phrase alignment in parallel treebanks. In Jan Hajic and Joakim Nivre, editors, *Proc. of the Fifth Workshop on Treebanks and Linguistic Theories*, pages 91–102, Prague, December 2006.

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

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