Qualitative and Quantitative Analysis of Annotators' Agreement in the Development of Cast3LB

M. Civit†, A. Ageno‡, B. Navarro∗, N. Bufí†, M.A. Martí†
†CLiC Centre de Llenguatge i Computació
Adolf Florensa s/n (Torre Florensa) 08028 Barcelona
{civit, nuria}@clic.fil.ub.es; amarti@fil.ub.es
‡TALP Research Centre (UPC)
Jordi Girona nº 3 08034 Barcelona
ageno@lsi.upc.es
∗ Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante Campus de San Vicente del Raspeig
Apartado 99. 03080 Alicante
borja@dlsi.ua.es

1 Introduction

The main objective of this paper is to present a qualitative and quantitative analysis of disagreement among annotators in the development of the syntactic annotation of Cast3LB corpus. Nowadays, this corpus is under development in the 3LB project[1], and includes 100.000 Spanish words. At syntactic level, more than 75% of the corpus has already been annotated.

According to [12], there are three main issues in the development of Treebanks: a) *systems for deriving structure automatically from unannotated language samples - parsers;* b) *specification of schemes of annotation-targets for parser output;* c) *metrics for quantifying parsing accuracy.*

From a general point of view, these metrics[2] measure the accuracy of an analysis regarding a preestablished *gold-standard*. They have been used to compare different analysis systems with the same reference corpus. The objective of these

[1]Project supported by Spanish Government, Ministerio de Ciencia y Tecnología, PROFIT program (FIT-150500-2002-244). This work has been partially funded by the X-TRACT-II project (BFF2002-04226-C03-03), too.

[2]The first definitions of these metrics appear at the PARSEVAL workshop.

metrics is to provide information about the similarity of the data, but they do not provide information about the location of disagreement into the analysis nor about its nature.

The linguistic annotation of corpora is a complex task. On the one hand, some linguistic expressions appearing in corpora show problems that do not appear in grammars (or do not appear explicitly enough). On the other hand, it is possible to give different syntactic analyses to a given linguistic structure, all of them being correct. Finally, each person has his/her particular view of the language: each one interprets sentences in a subjective and specific way.

Since the annotation process is a teamwork, it is important to know and assess the degree of consistency among the analyses given by each annotator. The annotation consistency is necessary to increase the quality of the corpus and to increase the utility of the corpus for the training of automatic systems or for linguistic research.

Another important issue to which very little attention has been paid to date is the evaluation of the accuracy among annotators. The question is: *How precisely can human beings analyze language structure?* [12].

There is a limit of human ability to analyze the own language, so there is a limit in the accuracy of human annotation. Deep studies about the annotators' consistency are rare[3]. However, we can point out the work carried out in the NEGRA project [5]. Nowadays, G. Sampson and A. Babarczy are working in an experiment in order to compare *the output of two human analysts applying the same parsing scheme independently to the same language samples* [12]. They use the SUSANNE scheme to annotate a set of 20.000 words from the BNC. The quality measure of a Treebank depends on the degree of agreement among annotators, whether there are errors or not. As it seems that errors are unavoidable, the main goal is to reduce as much as possible annotators' disagreement. But we have to take into account that even if there is a limit in human performance, this performance is the upper-bound for automatic language analysis. Indeed, the way in which humans solve these problems is the reference criteria for the automatic analysis of languages.

In this work, we present a study about annotators' agreement building the Cast3LB Treebank.

The first objective is to study the quantitative agreement among annotators at the constituent annotation level (see section 3). From the quantitative agreement, we obtain measures about the annotators' agreement, so we get the consistency of the syntactic annotation. Results over 90% are good enough to consider the

[3]There are some analyses of the accuracy in the semantic and morphological annotation [13] and [1].

resulting annotated corpus as a good resource for syntactic analysis. The second objective is to study the qualitative agreement among annotators. From the qualitative agreement, we want to analyze and classify the specific cases of disagreement. At this point, we follow the proposal of Sampson and Babarczy [12].

Section 2 presents the methodology followed in this study. Section 3 presents the main characteristics of the Cast3LB project. In sections 4 and 5, the quantitative and qualitative analysis are exposed. Finally, some conclusions are given in section 6.

2 Methodology

The purpose of this work is to make some contributions to the definition of a methodology for building treebanks. The main steps we have followed are:
1.- definition of the main principles of the syntactic annotation;
2.- annotation of a subset of 100 sentences according to these principles;
3.- enlargement of the guidelines in order to increase its coverage;
4.- annotation of 220 sentences and refinement of the guidelines;
5.- checking of these 320 sentences (steps 2 and 4) against the annotation guidelines, discussion and redefinition of the guidelines;
6.- annotation of 650 new sentences following the new guidelines;
7.- annotation of the last 30 sentences in order to perform the qualitative evaluation.

In every step, an automatic evaluation of the quantitative agreement has been carried out.

3 The Cast3LB Project

The Cast3LB project is the Spanish part of the 3LB project[4]. The objective of 3LB is to build three corpora linguistically annotated: one for Catalan (Cat3LB), one for Basque (Eus3LB) and one for Spanish (Cast 3LB)[5].

At the syntactic annotation level, we have followed two steps: the first is to bracket and tag the main constituents of each sentence; the second is to assign a function tag to the main constituents of each sentence.

The main points of the annotation scheme are[6]:

- Only explicit elements of the sentences are annotated. However, since we annotate anaphoric and coreferential relations, we have decided to introduce

[4]URL: *http://www.dlsi.ua.es/projectes/3lb/index_en.html*
[5]See [11] for more details about the composition of the corpus, the annotation levels, etc.
[6]These principles are the same than the Catalan corpus Cat3LB.

a special node for elliptical subjects. Regarding the verbal ellipsis, we mark this linguistic phenomenon with the symbol (*) in the sentence tag.

- We do not alter the word order. Spanish is a free constituent order language and this order has functional and communicative contents. If we change this order, we lose this information and alter the original data.
- We follow the constituency annotation.
- Finally, our aim is to develop a *neutral* annotation scheme, in the sense that we do not follow any specific linguistic theory. Our objective is to develop a linguistic annotated corpus useful for as many people as possible, so, if we follow a specific linguistic theory, the result of the project will be close to this theory and may become unuseful for some studies or purposes. Our idea is to give an *unmarked* annotation with respect to any theory.

The general annotation scheme of Cast3lb is described in [8] and [9].

The corpus has previously been morphologically annotated and disambiguated, on the one hand, and automatically parsed with a chunker [7], on the other. The work of human annotators is focused on the bracketing and labelling of each constituent.

In order to facilitate the annotators' task, we have adapted and developed some annotation tools: we are using an adaptation of the AGTK-toolkit [10] to do the syntactic annotation and will use 3LB-SAT [3] to do the semantic one.

4 Quantitative Analysis

In this section we present the quantitative analysis of discrepancies among annotators. Firstly, we describe the measures used to do so; then, we present the results, and finally we discuss them.

4.1 Description

Since no specific measures for the quantitative comparison of the annotator's agreement exist, we have decided to use some of the measures used for the evaluation of grammars and/or analysis methods. The need of an accurate evaluation when developing wide coverage analysers has been plainly agreed upon. It is out of the reach of this paper to describe in detail the existing evaluation systems (as an example, two excellent reviews of the different methods defined starting from 1991, [6] and [2] can be consulted). In our case, we have decided to use what might be considered the first objective measures, namely the ones defined in the Parseval

workshops [4], originally in order to evaluate syntactic wide-coverage analysers for the English language. Though not exclusive, its use is quite standardised for the evaluation of grammars and/or analysis methods, comparing the similarity of the results obtained with the reference parse trees (the ones previously considered *correct*, also known as *gold standard*). These similarity measures are based on the comparison of the constituents of both parse trees, on the ground of their spanning (their initial and final position in the sentence) as well as of their label, using recall (an attempt to measure coverage) and precision (a standard measure of accuracy). The specific metrics used are the following ones:

1. *Labelled Precision Rate*: number of constituents in the evaluated parse tree that coincide completely (both label and spanning) with one constituent in the reference parse, divided by the total number of constituents in the evaluated parse tree.

2. *Bracketed Precision Rate*: number of constituents in the evaluated parse tree whose spanning coincides with that of any constituent in the reference parse, divided by the total number of constituents of the evaluated parse tree.

3. *Labelled Recall Rate*: number of constituents in the evaluated parse tree that coincide completely (both label and spanning) with one constituent in the reference parse, divided by the total number of constituents in the reference parse tree.

4. *Bracketed Recall Rate*: number of constituents in the evaluated parse tree that span the same as some constituent in the reference parse, divided by the total number of constituents of the reference parse tree.

5. *Consistent Brackets Recall Rate*: number of constituents of the evaluated parse tree not crossing with any constituent in the reference parse tree, divided by the total number of constituents of the reference parse tree. It is considered that a constituent whose boundaries are $[i, j]$ crosses with another constituent with boundaries $[i', j']$ iff $i < i' <= j < j'$, that is, if the boundaries overlap but no constituent is completely included in the other one.

In other words, *recall* indicates the proportion of correct constituents that are hypothesized, whereas *precision* is the portion of hypothesized constituents that are correct. Also, the two bracketed measures are less strict, since they only regard those words of the sentence which are spanned by the constituents, ignoring the nonterminal label assigned to them. As to the Consistent Brackets Recall Rate, this measure is even less strict, since it considers only the proportion of constituents of

the evaluated parse tree which are inconsistent with the reference parse tree, that is, that could never be in the same parse tree.

However, it must be taken into account that, in our case, we are not evaluating the annotation performed by a certain analysis method, but comparing the annotations carried out by two linguists. Thus, neither of the analysis being compared can be considered the reference one, a *gold standard* does not exist. That's why we have decided to firstly compare both analysis in both senses (the analysis obtained by the first linguist with the analysis obtained by the second one, and then the other way round), and then to regard both measures in order to compute the averages. Considering the definitions of the measures described above, this fact implies that, somehow, the concepts of precision and recall do not make sense anymore, being unified in an only comparative measure (which we will denote precision, either labelled (LP) or bracketed (BP)).

The quantitative evaluation of the agreement has been performed in five steps, along which some of the disagreement problems have been progressively solved:

1. In the first step, once the basic annotation principles had been established, 100 sentences were annotated and criteria revised and enlarged. In this step, the annotation process of each sentence took about 20 minutes (as the process was completely manual).

2. In the second step, 220 sentences more were annotated. A first version of the guidelines which included more details about the adopted system arose from the discussions on the annotation schema. Here, an annotator needed 1 hour to annotate 5 sentences.

3. In the third step, the previous annotations were reviewed and compared so as to check both whether the guidelines did not contain any ambiguities and whether the annotators were already familiar with the adopted working system. Since the third step was a revision of the previously annotated sentences, the average here was about 14 sentences per hour.

4. In the fourth step, 670 sentences more were annotated. The average time spent was 9 sentences per hour.

5. The fifth step corresponds to the results of the experiment of the annotation evaluation over the last 30 sentences described in section 5. This final step took three hours (10 sentences per hour).

4.2 Results

Table 1 shows the results of the quantitative analyses (*LP* stands for *labelled precision rate*; *BP*, for *bracketed precision rate*; and *CB* for *Consistent brackets recall rate*).

	LP	**BP**	**CB**
Step 1	0.63359	0.72611	0.81072
Step 2	0.71166	0.80454	0.87124
Step 3	0.76537	0.84762	0.90487
Step 4	0.79222	0.85979	0.90821
Step 5	0.86927	0.90889	0.94958
same-lenght Sentences			
Step 3	0.85672	0.91683	0.95485
Step 4	0.90155	0.93323	0.96034
Step 5	0.91529	0.94036	0.96985

Table 1: Quantitative Evaluation

Logically, the increase in the measures is less pronounced as the steps advance, except for a significant increase from step four to step five. Besides, it can be observed that the bracketed precision gets to improve almost a 27% from the initial step to the final one (from .72 to .90). The bracketed precision is improved in more than 20% (from .63 to .86), and the consistent brackets recall rate in near 15%, from .81 to .94 (obviously, the less strict the measure is, the more difficult the possible improvement becomes).

4.3 Discussion

One of the main sources of disagreement among the annotators, which arose during the first stages of the analysis, was whether to consider as a single word certain complex structures such as *desde que* ('since'), *dar lugar a* ('give rise to'), *a lo largo de* ('along'), etc. Annotators adopted different criteria when labelling and bracketing these units. This affected the lenght of the sentences, for if such expressions were taken as multi-words, there were fewer terminal elements (words) in the sentence than if they were taken separately. Since our measures take into account the starting and finishing points of each constituent in the sentence, the fact that the length of the sentence varied implied a substantial decrease of the measures (and the closer to the beginning of the sentence these differently considered multi-words were, the higher this decrease of the measures would be). This issue has been accurately analysed and very strict criteria to deal with multi-words have

now been stablished in the guidelines. However, our aim has been to evaluate also the agreement figures obtained only for those sentences whose lengths coincided. Table 1 shows all the results obtained, including the evaluation of the measures for this mentioned subset (in this case only from the third annotation step, which was the time when this important source of disagreement was detected). If we just consider this subset of parse trees, it can be observed that labelled precision gets to improve above 30%, bracketed precision almost a 23%, and consistent brackets recall rate improves by 16%. As a result, bracketed precision raises to .94 and labelled precision to .91, and all the final values are comfortably over the 90% of agreement, probably getting closer to that limit in the precision of the annotation we mentioned in the introduction.

Figure 1 shows the evolution of the measures along these five steps.

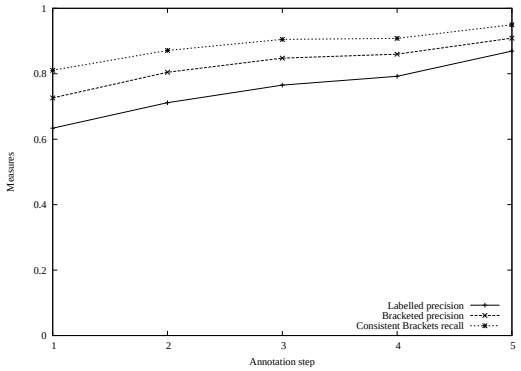

Figure 1: Evolution of annotators' agreement

5 Qualitative Analysis

In this section we present the qualitative analysis of discrepancies among annotators. Firstly, we describe the methodology; then we present the results, and finally we discuss them.

5.1 Description

To carry out the qualitative analysis of the discrepancies among annotators, five annotators were asked to annotate 30 sentences (1038 words; 31,45 words per sentence) of a randomly selected text about chemistry. Then, the results were manually compared one to one in order to find and classify the discrepancies. The classification has been made according to the typology suggested by [12]:
Type 1 *The language is inherently unclear/ambiguous.*

Type 2 *The language is clear but the guidelines are vague/missing/contradictory; it would be possible to extend the guidelines to give a predictable analysis in such cases.*
Type 3 *As 2, but it would be difficult to devise a suitable extension to the guidelines and handle such cases.*
Type 4 *The language and the guidelines are unambiguous, but one or bothannotators failed to apply the guidelines correctly.*

5.2 Results

As it can be seen in table 2, the main source of discrepancies are annotators' errors when applying the guidelines (type 4), and language ambiguities (type 1), whilst types 2 and 3 (discrepancies due to lacks in the guidelines) account for only a 15% of the differencies.

Type 1	**Type 2**	**Type 3**	**Type 4**
25.74%	12.17%	2.39%	59.86%

Table 2: Classification of discrepancies

Type-1 includes mainly attachment ambiguities of prepositional phrases and relative clauses as well as attachment ambiguities due to coordination phenomena.

The infinitive clause [7] *observar conducción de contacto entre los átomos metálicos* illustrates the first case of ambiguity: the last prepositional phrase (*entre los átomos metálicos*) could be attached to the previous noun (*contacto*), to the noun *conducción* or to the clause. According to the guidelines, the nesting node should be the clause, because we defined a default attachment: the highest node to the left. However, only two annotators gave the *correct* solution, while two did the nearest attach and another one gave the other possibility.

An ambiguous attachment related to a case of coordination is the next one[8]: *fibrillas o partículas metálicas* in which the adjective after the two coordinated nouns can refer to the nearest one or to both. As in the previous case, and according to the guidelines, the adjective should be annotated as a modifier of the coordinated node. Nevertheless, two of the annotators attached the adjective to the nearest noun.

[7] Lit. *to observe conduction of contact among the atoms metallic*; Trans. *to observe contact conduction among metal atoms*.

[8] Lit: *small_fibers or particles metallic*; Trans. *small fibers or metal particles*. There is no ambiguity in English because of the position of the adjective.

In type-2, the linguistic structure is clear, but the guidelines do not specify how some constituents should be labelled nor where some elements should be located. What we consider in this type of discrepancy are items general enough to be included in the guidelines, like the place of punctuation marks or enumerated lists. An example could be a comma before a coordinating conjunction; the guidelines did not specify whether it should be attached to the coordinated element or to the node containing the conjunction. This happens, for instance, in the next sequence[9]: *los metalomacrociclos , es_decir , complejos metálicos* for which there were no criteria in the guidelines. In all these cases, the guidelines were enriched with this information. As for commas, the decision was that coordination nodes should only contain the coordinating clause, so that commas belong to the previous constituent.

Type-3 refers to particular phenomena which appear in domain specific texts containing concrete and unfrequent structures that one cannot expect to find in guidelines conceived to annotate general/unrestricted text. This is the case of mathematical formulae or some peculiar conventions of the domain, like the following: *poli_(_3_-_alquiltiofenos_)* whose previous segmentation was *ncms000_ poli, Fpa_(, z_3, Fg_-, ncmp000_alquiltiofenos, Fpt_)*

Finally, type-4 includes errors such as oversights of the annotators: forgetting the suffix **.co** in the coordination label, giving one tag instead of another (S.F.ACons instead of S.F.AConc[10].), in spite of the annotation system which only allows to use pre-established tags.

5.3 Discussion

With type-1 discrepancies the problem is the perception of the ambiguity. If the annotator is able to find out two or more interpretations to be given to one structure, then he or she should know how to solve the problem (the highest attach to the left) from the guidelines; but if he or she does not realize that a structure is ambiguous, then he or she proposes a (correct) analysis according to his/her interpretation. What could be done is to analyze why an annotator performs one attachment or another (he or she has realized the ambiguity and so follows the guidelines; in spite of having realized the ambiguity, he or she makes a mistake; he or she proposes the only possible analysis according to the his/her interpretation of the structure, etc.).

Only 14% of discrepacies are due to lacks in the guidelines, which means that its coverage is large enough, even if the complete guidelines will be only available at the end of the project. One way to help annotators to follow the guidelines is

[9] Lit: *the metallomacrocycles, that is, complexes metallic*; Trans: *metallomacrocycles, that is, metall complexes*.

[10] *S.F.ACons* stands for consecutive adverbial subordinated clauses and *S.F.AConc* for concessive ones.

to make them redundant and clearer. On the one hand, if one wants information about noun phrase adjunction, for instance, the topic can be dealt with in the noun phrase section or in the adjunction one; on the other, the table of contents may not be detailed enough; thus, with clearer (well-structured or redundant) guidelines, the annotators' task could be facilitated.

Finally, even if type-4 discrepancies are by far the commonest (60%), in real terms they are the source of only 4% of discrepancies.

On the other hand, both types 1 and 4 are unavoidable in the process of annotation, while types 2 and 3 can be progressively reduced by means of the enrichment of the guidelines.

6 Conclusion

In this paper we have presented a methodology for evaluating the degree of qualitative and quantitative agreement among annotators during the progressive construction of the treebank. The results show a significant improvement from the first to the last step. The evaluation has been done from a quantitative as well as a qualitative point of view. The achieved degree of confidence makes this treebank well-suited to be used as a gold-standard because the degree of consistence/coherence is high enough. As a result, the developed guidelines ensure (almost) all the cases that can be found in our corpus and represent the basis for further developments of a full grammar for Spanish. Finally, the qualitative evaluation of the discrepancies is a key point in the annotation of treebanks that, up to now, has not been deeply examined.

Only a qualitative analysis of disagreements will allow to improve annotation methods, as the more difficult aspects of this process could be improved.

References

[1] A. Babarczy, J. Carroll and G. Sampson (2001), "Annotator error rates for part-of-speech tagging", *LINC2001, at 34th SLE*, Leuven.

[2] S. Bangalore, A. Sarkar,C. Doran and B.A. Hockey (1998), "Grammar & Parser Evaluation in the XTAG Project", *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, Granada, Spain.

[3] E. Bisbal,A. Molina, L. Moreno, F. Pla, M. Saiz-Noeda and E. Sanchís, (2003) "3LB-SAT: Una herramienta de anotación semántica", *Procesamiento del Lenguaje Natural*, n. 31, pp: 193-99, Alcalá de Henares

[4] E. Black, S. Abney,D. Flickinger,C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini and T. Strzalkowski (1991), "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars", *Proceedings of the Speech and Natural Language Workshop*, pp: 306-311, Pacific Grove, CA. DARPA.

[5] T. Brants (2000), "Inter_Annotator Agreement for a German Newspaper Corpus", *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

[6] J. Carroll, T. Briscoe and A. Sanfilippo (1998), "Parser Evaluation: a Survey and a New proposal", *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, Granada, Spain.

[7] M. Civit (2003), *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*, PhD. Thesis, Universitat de Barcelona.

[8] M. Civit and M.A. Martí (2002) "Design Principles for a Spanish Treebank", *Proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002)*, Sozopol, Bulgaria.

[9] M. Civit, M.A. Martí, B.Navarro, N. Bufí, B. Fernández and R. Marcos (2003), "Issues in the Syntactic Annotation of Cast3LB", *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC03), Workshop of the 10th EACL Conference*, Budapest, Hungary.

[10] S. Cotton and S. Bird (2000), "An integrated Framework for Treebanks and Multilayer Annotations", *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

[11] B. Navarro, M. Civit, M.A. Martí, B. Fernández and R. Marcos (2003), "Syntactic, semantic and pragmatic annotation in Cast3LB", *Proceedings of the Corpus Linguistics*, Lancaster, UK.

[12] G. Sampson and A. Babarczy (2003), "Limits to annotation precision", *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC03), Workshop of the 10th EACL Conference*, Budapest, Hungary.

[13] J. Véronis (2000), "Sense Tagging: don't look for the meaning but for the use", *Computational Lexicography and Multimedia Dictionaries, COMLEX*, Kato Achia, Greece.