Abstract. In this paper we present the conversion of two treebanks (Cat3LB for Catalan, and Cast3LB for Spanish) from its original constituent format into dependencies. The process has been done automatically but by manually writing the head and the function table. The process has also been used to improve the quality of the first annotation and to modify the annotation for further extensions of the treebanks. Treebanks in both formats are freely available for research purposes.

1 Introduction

In this paper we aim at presenting the conversion of two constituent treebanks into dependency ones. On the one hand, it is commonly accepted that constituent annotation is richer than that of dependencies since it contains different descriptive levels having a wide range of variability in the internal structure of constituents. Furthermore, in this kind of annotation, the head of each constituent can be easily inferred from the information contained in the constituents. On the other hand, dependencies provide an immediate description, without intermediate descriptive levels, because each tree node corresponds to a word. Therefore, it is easier to go from a constituent structure to a dependency one: heads can be easily obtained and intermediate levels can be avoided so as to obtain a complete standard dependency representation. Whereas, when going from a dependency structure to a constituent one, the result is a quite flat constituent structure lacking intermediate description levels. According to [10] and [11], dependency trees allow for more meaningful error measures and comparisons; and further works have confirmed this idea (see [1]).

Our starting point have been two constituent treebanks, one for Catalan and another for Spanish, which have been converted into the dependency format. This way, two corpora in two different formats are made available to the community, with the hope on enlarging the research on NLP for the two concerned languages.

1 The work described here has been partially funded by the CESS-ECE project (HUM2004-21127).

© Springer-Verlag Berlin Heidelberg 2006
In section 2 we present the treebanks from which the conversion has been done; in sections 3 and 4 we discuss about the head table and the function table used for the conversion; in section 5 we deal with constituent tags. In section 6 we discuss about the conversion process and its influence for the improvement of the annotation in the original treebank. Last section, section 7, is devoted to conclusions and further work.

2 Starting Point: 3LB Treebanks

The Spanish treebank, Cast3LB and the Catalan one, Cat3LB, consist of 100,000 words each (approximately 4,000 sentences for Cast3LB and 2,800 for Cat3LB). For both treebanks a theory-neutral and surface-oriented annotation scheme has been adopted, which follows the linguistic tradition for Romance Languages. There are three levels of annotation: morphology (pos-tag plus lemma), constituency (tree structure with non-terminal and terminal nodes); and functions, although these are only given to sentence structure nodes. No nodes have been added to the trees, except those for elliptical subjects (Spanish and Catalan being pro-drop languages). Figure 1 shows the analysis of the Spanish sentence (Quiero con esto decir que Medardo Fraile ha escrito un relato extraño y divertido) in the constituency format.

Fig. 1. Constituent format treebank

One of the main problems to be dealt with during the annotation process is discontinuity. In the case of Cat3LB and Cast3LB indexes have been used as suffixes for constituents and function tags when constituents are discontinuous. There are two cases of discontinuity made explicit in Cast3LB and Cat3LB: one related to constituents and the other related to functions. The first one


3 Word-by-word translation: Want with this to say that Medardo Fraile has written a story strange and funny; Translation: With this, I mean that Medardo Fraile wrote a strange, funny story.