Design and Development of a System for the Detection of Agreement Errors in Basque

Arantza Díaz de Ilarraza, Koldo Gojenola and Maite Oronoz

Department of Computer Languages and Systems
University of the Basque Country
P.O. box 649, E-20080 Donostia
{jipdisaa, jibgogak, jiporanm}@si.ehu.es

Abstract. This paper presents the design and development of a system for the detection and correction of syntactic errors in free texts. The system is composed of three main modules: a) a robust syntactic analyser, b) a compiler that will translate error processing rules, and c) a module that coordinates the results of the analyser, applying different combinations of the already compiled error rules. The use of the syntactic analyser (a) and the rule processor (b) is independent and not necessarily sequential. The specification language used for the description of the error detection/correction rules is abstract, general, declarative, and based on linguistic information.

1 Introduction

The problem of the detection and correction of syntactic errors has been addressed since the early years of natural language processing. Different techniques (Vandeventer, 2003) have been proposed for the treatment of the significant portion of errors (typographic, phonetic, cognitive and grammatical) that result in valid words (Kukich, 1992). Although many commercial grammar checkers have been developed (Paggio and Music, 1998), there is little published work on their implementation or evaluation. This is due in part to the fact that the mechanisms used for the implementation have not been very sophisticated (as in some systems that use a large set of regular expressions) and also that commercial companies are not willing to reveal implementation details about their tools. The aim of the present work is to examine the feasibility of corpus-based syntactic error detection focusing in detecting agreement errors.

The system will be applied to Basque, an agglutinative language with relative free order among sentence elements. In our research group, work in error detection at morphological level has already been accomplished and a spelling checker-corrector (Aldezabal *et al.* 1999), - called XUXEN - was marketed 10 years ago. Error detection at syntactic level needs of a robust syntactic analyser and we will use the Basque syntactic analyser (Aduriz and Díaz de Ilarraza, 2003) that was developed using Constraint Grammar (CG) (Karlsson *et al.*, 1995).

Figure 1 shows the syntactic analysis chain in which sequential rule layers, most of them materialised in Constraint Grammars, enrich the output of the previous layer with the respective information.