Constructing a Parser for Latin

C.H.A. Koster

Computing Science Institute,
University of Nijmegen,
The Netherlands,
E-mail: kees@cs.kun.nl

Abstract. We describe the construction of a grammar and lexicon for Latin in the AGFL formalism, in particular the generation of the lexicon by means of transduction and the description of the syntax using the Free Word Order operator. From these two components, an efficient Top-Down chart parser is generated automatically. We measure the lexical and syntactical coverage of the parser and describe how to increase it.
The morphological generation technique described here is applicable to many highly-inflected languages. Since the Free Word Order operator described can cope with the extremely free word order in Latin, it may well be used for the description of free-word-order phenomena in modern languages.

1 Introduction

Why would anybody in his right mind construct a formal grammar of Latin? Although there exist some active speakers of the language and according to some its best poetry was produced in the nineteenth century, the language is as dead as a doornail. A large corpus of latin texts is extant, but there is no expectation of any important additions. Most texts have been translated into many other languages in which they can be enjoyed without the drudge of learning Latin. Commercial application of an Information Retrieval system for Latin is inconceivable. Furthermore there already exists an overwhelming number of learned grammars and dictionaries for it, to which any formal grammar would add nothing new.

It was for the Latin language, and earlier for Greek, that the science of linguistics as we know it was developed. Everyday concepts and terminology of Latin still pervade western linguistic thinking. The understanding of the structure of Latin provides a framework in which not only its linguistic relatives, but also utterly unrelated languages could be analysed, modeled and described. The Latin language has a number of properties (detailed and rich morphology, very free word order) which together with its quite regular structure make it an interesting object for formal description. Practically, it is the mother of the Romance languages and the aunt of many other languages (including English) which do have practical and even commercial value. Lastly, describing it with the aid of modern grammar and parsing technology may be of therapeutic value for one who has been forcefed on it for a number of years in high school.