Certified Parsing

Background

Parsing is the act of transforming plain text into some structure that can be analysed by computers for further processing. One might think that parsing has been studied to death and after *yacc* and *lex* no new results can be obtained in this area. However recent results and novel approaches make it increasingly clear, that this is not true anymore.

We propose to approach the subject of parsing from a certification point of view. Parsers are increasingly part of certified compilers, like *CompCert*, which are guaranteed to be correct and bug-free. Such certified compilers are crucial in areas where software just cannot fail. However, so far the parsers of these compilers have been left out of the certification. This is because parsing algorithms are often adhoc and their semantics is not clearly specified. Unfortunately, this means parsers can harbour errors that potentially invalidate the whole certification and correctness of the compiler. In this project, we like to change that.

Only in the last few years, theorem provers have become good enough for establishing the correctness of some standard lexing and parsing algorithms. For this, the algorithms need to be formulated in way so that it is easy to reason about them. In earlier work about lexing and regular languages, the authors showed that this precludes well-known algorithms working over graphs. However regular languages can be formulated and reasoned about entirely in terms regular expressions, which can be easily represented in theorem provers. This work uses the device of derivatives of regular expressions. We like to extend this device to parsers and grammars. The aim is to come up with elegant and useful parsing algorithms whose correctness and the absence of bugs can be certified in a theorem prover.

Proposed Work

One new development in formal grammar is the Parsing Expression Grammar (PEG) which is proposed as an refinement of standard Context Free Grammar. The aim of this extension is to internalize disambiguition normally done with semantic methods. The idea is to introduce negative, conjunctive operators as well as production priorities, so that the grammars written in PEG are unambiguous in the first place. Another benefit of PEG is that it admits a very efficient linear parsing algorithm.

However, one disadvantage of PEG is that it does not allow left recursion in grammar specification, i.e., standard parsing algorithms of PEG can not deal with left recursion. Although some authors claimed PEG parsing algorithms for left recursion, none of them provide correctness proof, not even in paper-and-pencil form.