Multilingual Support in Bison-based Parser

Background

Programming and scripting languages provide a common way to describe complex behavior into software. For example, most DBMSs implement SQL to manipulate databases. The input to software is first syntactically analyzed by a parser. A popular tool to implement a parser of a large complex language is Bison, since it allows writing complex grammar using BNF-like notation and demonstrates good performance. Bison reads input through a lexer, which scans input for lexical tokens. The most commonly used lexer together with Bison is Flex. Unfortunately, Flex is designed to work with 8-bit characters, i.e., for ASCII encoding, and thus badly support modern requirements to process programs in multilingual encodings such as UTF-8 and UTF-16.

Purpose and Scope

The purpose of this project is to find a state-of-the-art lexer and implement a prototype of a parser combing the chosen lexer with Bison for a SQL-based language.

The following tasks are included in the project:

- Analyze requirements for a modern parser
- Research the state-of-the-art in lexers
- Implement scanner for a Bison parser supporting SQL-based language
- Analyze performance of the implementation

The project is intended for one or two students.

Experience and Knowledge requirements

- Deep program development experience
- Experience with parsing technologies and, e.g., with YACC or Bison
- Experience in C/C++
- Experience with multilingual encodings
- Experience from using SQL
- Knowledge needed to perform thesis work

Presentation of results

The project should result in a working prototype and a report including the prerequisites, assumptions, individual performance result and conclusions.

Contacts for application and questions:

Ruslan Fomkin, **Ruslan.Fomkin@starcounter.com,** 073 – 059 5789