<https://github.com/Horatzioscool/formal-languages-and-compiler-design>

The above repository contains the source code for a scanning algorithm for my proposed programming language: Englishlike

The source code contains interface and class definitions for a Scanner, Symbol Table and Program Internal form.

**Scanner**: The scanner has a single exposed method, ‘Scan’ which receives a StreamReader.

The scanner contains a Symbol Table and a Program Internal Form.

During the execution of the ‘Scan’ method the scanner will mutate the Symbol Table and Program Internal Form by adding parsed tokens to them.

The detection algorithm of the scanner will continue reading until a separator or a delimiter is reached. It then returns as a token all the non-separator, non-delimiter input it found until then.

Delimiters are returned as a single token, separators are bunched together and returned as a token.

The scanner will then attempt to match the token with several possible token types.

The first token type is a ‘reserved’ token type, which represents a reserved word, an operator, or delimiters. Reserved words are listed within the specification.

The second token type is identifier, for which a regex is used to match a letter-starting, alphanumeric string.

The last token type is constant, for which a regex is used to match either a string representing a numeric integer or a string constant.

The numeric integer is either 0 or a non-zero digit starting number, with either + or – beforehand.

The string constant is an alphanumeric string between double quotes “.

Reserved words are added to the Pif with code 0.

Identifiers and constants are added to the symbol table.

Identifier and constants are added to the Pif with their code being their index within the symbol table.

**Symbol Table**: The symbol table exposes two methods, insert and search: both receiving a string token as parameter and both will return an integer tuple.

The symbol table is internally implemented using a hash table composed of linked lists.

Tokens have their hash code computed, which is the sum of the ascii codes of its characters.

Tokens are then inserted in the linked list at the index which is the modulo of their hash code and the table’s size. If there is an already existing element their the token will be added at the back of the list.

**Program Internal Form**: The program internal form is a wrapper over a list of pairs of string, string. It only exposes an ‘Add’ method to add a new pair to the pif.