Лексический анализатор

Определим следующие типы распознаваемых токенов:

1. Ключевые слова. Список слов: if, else, then, while, begin, end, do, const, for, function, var, to, downto, +, -, \*, /, =, <, <=, >, >=, <>, :=, and, or, not, (, ).
2. Константы. Выделим три типа констант: строковые, целочисленные, вещественные.
3. Идентификаторы. Представляют собой строку однозначно определяющий идентификатор. Например: “x”, “y” и так далее.

Лексический анализатор – программа, принимающая на вход текст (последовательность символов из алфавита) и разбивающая его на подстроки (лексемы) в соответствии с некоторым набором правил.

Далее определим основные классы, необходимые для реализации лексического анализатора.

***CLexer.*** Представляет собой класс лексического анализатора. Хранит строковую константу разделителей токенов (separators), текст программы, строку, хранящую текущий токен, а также позицию текущего читаемого символа в тексте паскаль-программы.

Класс имеет следующие методы: GetTokenType() - определяет тип текущего токена и возвращает умный указатель на созданный объект класса (CIdentToken/ CKeywordToken/CConstToken).

GetNextToken() - выделяет следующий токен, начиная с позиции currentPosition. После выделения токена возвращает умный указатель на него.

IsReal (string s) - проверка строки s на соответствие типу данных real. Для того, чтобы строка соответствовала типу данных real, необходимо, чтобы в числе присутствовала одна точка. Возвращает bool.

IsInteger (string s) - проверка строки s на соответствие типу данных Integer. Чтобы строка соответствовала типу данных integer, необходимо, чтобы в строке присутствовали исключительно цифры.

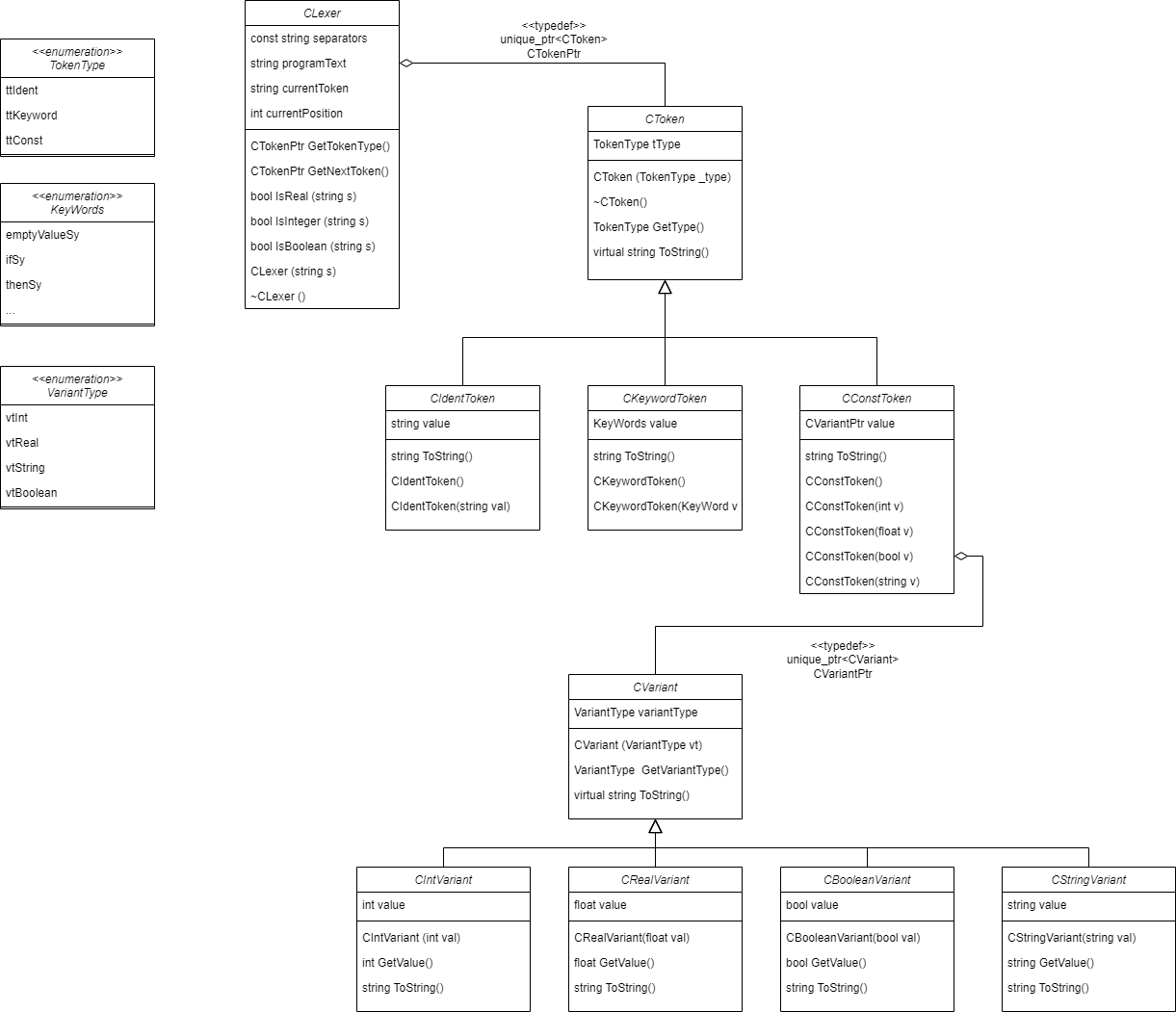
IsBoolean (string s) - проверка строки s на соответствие типу данных boolean. Строка s должна принимать одно из значений: “true”/”false”.

***CToken.*** Абстрактный класс. Наследуемые классы имеют поле типа TokenType, характеризующее один из трех видов токенов. Метод GetType() возвращает значение приватной переменной типа TokenType. Метод ToString() возвращает строковое представление токена.

***CIdentToken / CKeywordToken / CConstToken.*** Наследуются от CToken. Представляются как три разных вида токенов - идентификаторы, ключевые слова, константы. Все три класса переопределяют метод ToString(), а также имеют различные поля. CIdentToken имеет поле string, т.к. идентификатор представляет собой строку. CKeywordToken имеет поле KeyWords, где хранит enum значение одного из ключевых слов. CConstToken имеет поле CVariantPtr, предоставляющее умный указатель на объект класса CVariant, хранящий значение константы.

***CVariant.*** Абстрактный класс. Наследуемые классы имеют поля типа VariantType - enum значение, характеризующее тип константы (string/real/integer/boolean), а также методы GetVaraintType(), возвращающее тип константы и ToString(), преобразующий константу в строку.

***CIntVariant/CRealVariant/CBooleanVariant/CStringVariant.*** Наследуются от CVariant. Представляются как 4 разных вида констант - целочисленные, вещественные, строковые, логические. В каждом классе есть метод GetValue(), возвращающий значение константы, а также переопределенный метод ToString(), возвращающий значение константы в виде строки.



Умные указатели. В качестве указателей на объекты СToken и CVariant использовались умные указатели. Благодаря умным указателям возможно избежать утечек памяти в программе и обеспечить безопасное ее использование.