An implementation of a parsing algorithm for LALR grammars.

Ali Mousa Jaber
AN IMPLEMENTATION OF A PARSING ALGORITHM

FOR LALR GRAMMARS

by

ALI MOUSA JABER

A Thesis

Presented to the Graduate Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Computing Science

Lehigh University

1983
Certificate of Approval

This thesis is accepted and approved in partial fulfillment of the requirements for the degree of Master of Science.

May 2, 1983
(date)

Professor in Charge

Chairman of Department
Acknowledgment

The author wishes to thank Professor Samuel L. Gulden for his helpful suggestions and thoughts in preparation of this thesis.
# Table of Contents

1. ABSTRACT .................................................. 1

2. INTRODUCTION ............................................... 2

3. BACKGROUND ................................................ 5
   3.1 Basic Definitions and Notation .......................... 5
      3.1.1 Alphabets and Strings ............................ 5
      3.1.2 Terminals and Nonterminals ....................... 5
      3.1.3 Production Rules and Grammars ................... 6
      3.1.4 Languages ........................................ 7
   3.2 LR(k) Grammars .......................................... 7
      3.2.1 FIRST set ....................................... 7
      3.2.2 Eujgmented Grammars ............................. 9
      3.2.3 Definition of LR(k) Grammar ..................... 9

4. ANALYSIS OF THE PARSING ALGORITHM ....................... 11
   4.1 LR(0) Item ............................................. 11
   4.2 LR(1) Item ............................................. 12
   4.3 Sets of LR(1) Items .................................... 12
      4.3.1 ClDUSURE Function ................................ 12
      4.3.2 GDTO Function .................................... 13
   4.4 Sets of LALR(1) Items .................................. 14
   4.5 LALR(1) Parsing Table ................................ 15
   4.6 The Parsing Process .................................... 17

5. IMPLEMENTATION OF THE PARSING ALGORITHM ................ 19
   5.1 Input and FIRST set .................................... 19
      5.1.1 Procedure GETPROD ................................ 20
      5.1.2 Procedure FIRSTPROD ............................... 21
      5.1.3 Procedure FIND_LASET ............................. 23
   5.2 Sets of Items .......................................... 24
      5.2.1 procedure CLOSURE ................................ 25
      5.2.2 Procedure BUILD_NEUSTATE ......................... 27
   5.3 Parsing Table .......................................... 29
   5.4 Parsing Process ....................................... 31

6. CONCLUSIONS ................................................. 34

7. LIST OF REFERENCES ....................................... 35

8. VITA ...................................................... 36
1. ABSTRACT

AN IMPLEMENTATION OF A PARSING ALGORITHM FOR

LALR GRAMMARS

by Ali M. Jaber

An LALR parsing algorithm presented by Alfred V. Aho, and Jeffrey D. Ullman is presented in this paper. This class of languages is of great importance because the parsing tables obtained are considerably smaller than the LR tables, yet most common syntactic constructs of programming languages can be expressed conveniently by LALR grammar. An introduction about the LR parsing technique and a background on the theory of parsing are given. An analysis of the algorithm is shown.

Logically, the algorithm consists of two parts, the parsing table, and the driver routine. The parsing table is created dynamically using pointers. To implement the concept of the parsing algorithm, a program written in PASCAL is discussed.
2. INTRODUCTION

A grammar forms the underlying method in deriving sets of strings of a formal language. In this paper we shall deal with a special type of grammar known as an LALR grammar which is a subset of a particular class of grammars known as context free grammars (CFG), and we will present an LALR parsing algorithm.

LALR (lookahead-LR) grammars are a subset of LR grammars. An LR parser scans the input from left to right and constructs a rightmost derivation in reverse (hence, the name LR). The LALR parsing algorithm used in this paper is due to Aho, and Ullman.

Logically, an LR parser consists of two parts, a
The driver routine and a parsing table. The driver routine is the same for all LR parsers; only the parsing table changes from one parser to another. The parser has an input, a stack, and a parsing table as shown in Fig. 2.1. The input is read from left to right, one symbol at a time. The stack contains a string of the form $s \ X \ s \ldots$, where $s$ is on top. Each $X$ is a grammar symbol and each $s$ is a symbol called a state. The parsing table consists of two parts, a parsing action function $\text{ACTION}$ and a goto function $\text{GOTO}$.

The program driving the LR parser behaves as follows. It determines $s$, the state currently on top of the stack, and $a$, the current input symbol. It then consults $\text{ACTION}[s,a]$, the parsing action table entry for state $s$ and input $a$. The entry $\text{ACTION}[s,a]$ can have one of four values:

1. shift $s$
2. reduce $A \rightarrow \beta$
3. accept
4. error

The function $\text{GOTO}$ takes a state and a grammar symbol as argument and produces a state. It is essentially the transition table of a deterministic finite automaton whose input symbols are the terminals and nonterminals of
the grammar.

LR parsers can be constructed to recognize virtually all programming language constructs for which context free grammars can be written.
3. BACKGROUND

3.1 Basic Definitions and Notation

3.1.1 Alphabets and Strings

An alphabet is any set of symbols. An alphabet need not be finite or even countable, but for all practical applications our alphabets will be finite. The symbol $\sum$ is used to designate an alphabet.

A string is a sequence of elements drawn from an alphabet. For example, 01011 is a string over the binary alphabet $\{0,1\}$. The empty string is denoted by $\lambda$.

**DEFINITION** We formally define strings over an alphabet $\sum$ in the following manner:

1. $\lambda$ is a string over $\sum$.
2. If $x$ is a string over $\sum$ and $a$ is in $\sum$, then $xa$ is a string over $\sum$.
3. $y$ is a string over $\sum$ if and only if it is being so follows from (1) and (2).

3.1.2 Terminals and Nonterminals

A terminal is a member of $\sum$. In order to define structural rules for a grammar we introduce a finite set of objects called nonterminals. The set of nonterminals is disjoint from $\sum$, and is usually denoted by $N$. 
3.1.3 Production Rules and Grammars

A production rule, or production for short, has the general form \( X \rightarrow Y \) where \( X \) and \( Y \) are strings in the terminal and nonterminal sets of a given grammar. The \( Y \) may be an empty string, but \( X \) cannot be. Productions are used to generate strings in the language.

A grammar is a 4-tuple: \( G = (N, \Sigma, P, S) \), where:

- \( N \) is a finite nonempty set of nonterminals;
- \( \Sigma \) is a finite set of alphabet;
- \( P \) is a finite set of productions of the form \( X \rightarrow Y \), where \( X \) and \( Y \) are in \( (N \cup \Sigma^*) \), and \( X \) contains at least one element in \( N \);
- \( S \) is called the start symbol;

We represent a derivation step by the symbol \( \Rightarrow \), a sequence of one or more derivation steps by \( \Rightarrow^* \), and a sequence of zero or more derivation steps by \( \Rightarrow^* \).

**DEFINITION**  A context free grammar (CFG) is denoted by \( G = (N, \Sigma, P, S) \), where \( N, \Sigma, P \), and \( S \) as defined above; each production is of the form \( A \rightarrow W \), where \( A \) is a member of \( N \) and \( W \) is a string of symbols from \( (N \cup \Sigma)^* \).

An example of a context free grammar is given below:

\( N = \{E, T, F\}, \Sigma = \{+, *, (, ), a\}, S = E \), and \( P \) the set:
1. $E \longrightarrow E + T$
2. $E \longrightarrow T$
3. $T \longrightarrow T*F$
4. $T \longrightarrow F$
5. $F \longrightarrow (E)$
6. $F \longrightarrow a$

3.1.4 Languages

DEFINITION Let $G=(N, \Sigma, P, S)$ be a grammar. The language generated by $G$, written $L(G)$, is the set:

$L(G) = \{ w \in \Sigma^* \mid S \Rightarrow^* w \}$

Example: let $G=(N, \Sigma, P, S)$ where:

$N=\{A, C\}$, $\Sigma=\{c, d\}$, $S=A$, and $P$ is the set:

$A \longrightarrow CC$
$C \longrightarrow cC \setminus d$

Then, $L(G) = c^*dc^*d$.

3.2 LR(k) Grammars

Before we define the LR grammar, we need to define a useful function $\text{FIRST}(w)$.

3.2.1 FIRST set

The domain of $\text{FIRST}$ is some string $w$ in $\{N \cup \Sigma\}_k^*$. The function is defined as follows:
FIRST\( (W) = \{x \mid \text{w} \Rightarrow^* xy, x, y \in \sum^* \} \)

\[
\begin{align*}
|x| &= k \text{ if } y \neq \lambda \\
|x| &\leq k \text{ if } y = \lambda
\end{align*}
\]

where the derivations are left-most. That is, FIRST\( (W) \) for some string\( W \) is the set of all leading terminal strings of length \( k \) or less in the strings derivable from \( W \).

To compute the FIRST set we can use the following rules:

1. \( \text{FIRST} (aw) = a \text{FIRST}^k (W) \) for any string \( W \), \( a \in \sum \).
2. \( \text{FIRST} (\lambda) = \{\lambda\} \), \( \lambda \) being the empty string.
3. \( \text{FIRST} (XY) = \text{FIRST}^k (\text{FIRST}^k (X) \text{FIRST}^k (Y)) = \text{FIRST}^k (\text{FIRST}^k (X) \text{FIRST}^k (Y)) = \text{FIRST}^k (\text{FIRST}^k (X)Y) \)
   \( X, Y \in \{N \cup \sum\}^* \).
4. Given a production \( A \rightarrow^* W \) in \( G \), \( \text{FIRST} (A) \) contains \( \text{FIRST}^k (W) \), \( A \in N \).

Consider the example in section 3.1.3, then

\[
\begin{align*}
\text{FIRST} (E) &= \{(, a)\} \\
\text{FIRST} (T) &= \{(, a)\} \\
\text{FIRST} (F) &= \{(, a)\}
\end{align*}
\]
3.2.2 Augmented Grammars

**DEFINITION** Let \( G = (N, \Sigma, \mathcal{P}, S) \) be a CFG. We define the augmented grammar derived from \( G \) as \( G' = (N \cup \{S'\}, \Sigma, \mathcal{P} \cup \{S' \rightarrow S\}, S') \).

The augmented grammar \( G' \) is merely \( G \) with a new starting production \( S' \rightarrow S \), where \( S' \) is a new start symbol, not in \( N \). We assume \( S' \rightarrow S \) is the zeroth production in \( G' \) and that the other productions of \( G \) are numbered \( 1, 2, 3, \ldots, p \). We add the starting production so that when a reduction using the zeroth production is called for, we can interpret this "reduction" as a signal to accept.

3.2.3 Definition of LR(k) Grammar

**DEFINITION** Let \( G = (N, \Sigma, \mathcal{P}, S) \) be a CFG and let \( G' = (N', \Sigma, \mathcal{P}' , S') \) be its augmented grammar. We say that \( G \) is LR(k), \( k > 0 \), if the three conditions:

1. \( S' \xrightarrow{\beta} \alpha \leadsto \gamma \xrightarrow{\beta} \alpha \beta W \)

2. \( S' \xrightarrow{\beta} \gamma \beta \xrightarrow{\beta} \gamma \beta Y \), and

3. \( \text{FIRST}(W) = \text{FIRST}(Y) \)

imply that \( \alpha \beta \gamma = \gamma \beta \gamma \). (That is, \( \alpha = \gamma \), \( \beta = \beta \), and \( \gamma = \gamma \)). A grammar is LR if it is LR(k) for some \( k \).

Intuitively, this definition says that if \( \alpha \beta \gamma \) and \( \gamma \beta \gamma \)
\( \alpha \beta Y \) are right sentential forms of the augmented grammar with \( \text{FIRST}(W) = \text{FIRST}(Y) \) and if \( A \rightarrow^k \beta \) is the last production used to derive \( \alpha \beta W \) in a rightmost derivation, then \( A \rightarrow^k \beta \) must also be used to reduce \( \alpha \beta Y \) to \( \alpha AY \) in a right parse. Since \( A \) can derive \( \beta \) independently of \( W \), the \( LR(k) \) condition says that there is sufficient information in \( \text{FIRST}(W) \) to determine that \( \alpha \beta \) was derived from \( \alpha A \).
4. ANALYSIS OF THE PARSING ALGORITHM

4.1 LR(0) Item

DEFINITION An LR(0) item of a grammar G is a production of G with a dot at some position of the right side of the production. For example, production $A \rightarrow^* XYZ$ generates the four items:

- $A \rightarrow .XYZ$
- $A \rightarrow X.YZ$
- $A \rightarrow XY.Z$
- $A \rightarrow XYZ$.

The production $A \rightarrow^* \lambda$ (a empty string) generates only one item $A \rightarrow^* \lambda$.

Intuitively, an item indicates how much of a production we have seen at a given point in the parsing process. For example, the first item above would indicate that we are expecting to see a string derivable from XYZ next on the input. The second item would indicate that we have just seen on the input a string derivable from X and we next expect to see a string derivable from YZ.
4.2 LR(1) Item

**Definition**
An LR(1) item of a grammar G is an LR(0) item with a lookahead symbol. The general form of an item is [A ---> α·B , a], where A ---> α·B is a production and a is a terminal or the right end marker $\$$. A ---> α·B is called the core of the LR(1) item.

The lookahead has no effect in an item of the form [A ---> α·B , a], where B is not A, but an item of the form [A ---> α·, a] calls for reduction by A ---> α only if the next input symbol is a.

4.3 Sets of LR(1) Items

Let G' be the augmented grammar for G, to construct the sets of LR(1) items, we need to define two functions, CLOSURE and GOTO.

**4.3.1 CLOSURE Function**

If I is a set of items for a grammar G, then the set of items CLOSURE(I) is constructed from I by the rules:

1. Every item in I is in CLOSURE(I).
2. If [A ---> α·B , a] is in CLOSURE(I) and B ---> γ is a production, then add the item [B ---> γ , b] where b in FIRST(B) a), if it is not already there.

Intuitively, [A ---> α·B , a] in CLOSURE(I) indicates that, at some point in the parsing process, we next
expect to see a string derivable from $BB$ as input. If $B \rightarrow \gamma$ is a production, we would also expect to see a string derivable from $\gamma$ at this point. It is for this reason we also include $[B \rightarrow \gamma, b]$ in CLOSURE($I$). The procedure CLOSURE is shown in Fig. 4.1 below:

procedure CLOSURE($I$);
BEGIN
REPEAT
FOR each item $[A \rightarrow \alpha . B, a]$ in $I$, each production $B \rightarrow \gamma$, and each terminal $b$ in FIRST($\gamma$) such that $[B \rightarrow \gamma, b]$ is not in $I$ DO
add $[B \rightarrow \gamma, b]$ to $I$;
UNTIL no more items can be added to $I$;
RETURN $I$;
END;

Fig. 4.1

4.3.2 GOTO Function

The function GOTO($I, X$) where $I$ is a set of items and $X$ is a grammar symbol is defined to be the closure of the set of all items $[A \rightarrow \alpha . X . B, a]$ such that $[A \rightarrow \alpha . X . B, a]$ is in $I$. The procedure GOTO is shown in Fig. 4.2 below:

procedure GOTO($I, X$);
BEGIN
let $J$ be the set of items $[A \rightarrow \alpha . X . B, a]$, such that $[A \rightarrow \alpha . X . B, a]$ is in $I$;
RETURN CLOSURE($J$);
END;

Fig. 4.2

13
To construct the sets of items, we begin by computing the closure of \( \{S'\rightarrow\rightarrow S,S\} \). We match the item \( \{S'\rightarrow\rightarrow S,S\} \) with the item \( \{a\rightarrow\rightarrow \lambda, B, a\} \) in the procedure clousure. That is, \( A=S' \), \( K=\lambda \), \( B=S \), \( \beta=\lambda \), and \( a=s \). CLOUSURE tells us to add \( \{B\rightarrow\rightarrow \gamma, b\} \) for each production \( B\rightarrow\rightarrow \gamma \) and terminal \( b \) in FIRST(\( \beta a \)). The main procedure for constructing the sets of items is shown in Fig. 4.3 below:

```
procedure MAIN ;
   BEGIN
      C := \{CLOUSURE(\{S'\rightarrow\rightarrow S,S\})\} ;
      REPEAT
         FOR each set of items I in C and each grammar symbol X such that GOTO(I,X) is not empty and not already in C
         DO add GOTO(I,X) to C ;
      UNTIL no more sets of items can be added to C ;
   END ;
```

Fig. 4.3

4.4 Sets of LALR(1) Items
We are now prepared to give our LALR(1) sets of items construction algorithm. The general idea is to construct the sets of LR(1) items and merge the sets having the same core.

Algorithm 4.1
------------------
input : An augmented grammar \( G' \) for the grammar \( G \).
output: The sets of LALR(1) items.
method:

1. Construct \( C = \{ I_0, I_1, \ldots, I_n \} \), the collection of sets of LR(1) items.

2. For each core present among the sets of LR(1) items, find all sets having that core, and replace these sets by their union.

3. The GOTO is constructed as follows: If \( J \) is the union of one or more sets of LR(1) items, i.e., \( J = I_1 \cup I_2 \cup \ldots \cup I_m \), then the cores of \( GOTO(I_1, X), GOTO(I_2, X), \ldots, GOTO(I_m, X) \) are the same, since \( I_1, I_2, \ldots, I_m \) all have the same core. Let \( K \) be the union of all sets of items having the same core as \( GOTO(I_1, X) \), then \( GOTO(J, X) = K \).

4.5 LALR(1) Parsing Table

We now give the rules whereby the LALR(1) parsing action and goto functions are constructed from the sets of LALR(1) items.

ALGORITHM 4.2

Construction of a canonical LALR(1) parsing table:

**Input:** an augmented grammar \( G' \) for the grammar \( G \).

**Output:** If possible, the LALR(1) parsing table, i.e., the action function \( ACTION \) and goto function \( GOTO \).

**Method:**
1. Construct $C = \{ I_1, I_2, \ldots, I_n \}$, the collection of sets of LALR(1) items for $G$.

2. State $i$ of the parser is constructed from $I_i$, the parsing actions for state $i$ are determined as follows:

   - If $[A \rightarrow \alpha, a] \in I$ and $\text{GOTO}(I_i, a) = j$, then set $\text{ACTION}(i, a)$ to "shift $j$."
   - If $[A \rightarrow \alpha, A] \in I$, then set $\text{ACTION}(i, a)$ to "reduce $A \rightarrow \alpha$ ."
   - If $[S' \rightarrow S, \lambda] \in I$, then set $\text{ACTION}(i, \lambda)$ to "accept ."

If a conflict results from the above rules, the grammar is said not to be LALR(1), and the algorithm is said to fail.

3. The goto transitions for state $i$ are determined as follows: if $\text{GOTO}(I_i, A) = j$, then $\text{GOTO}(i, A) = j$.

4. All entries not defined by rules (2) through (3) are made "error."

5. The initial state of the parser is the one constructed from the set containing item $[S' \rightarrow S, \lambda]$.

The table formed by Algorithm 4.2 is called the LALR(1) parsing table. If there is no parsing action conflict, then the given grammar $G$ is called an LALR(1) grammar.
4.6 The Parsing Process

As mentioned earlier, the parser has an input, a stack, and a parsing table. The parsing table consists of two parts, a parsing action function ACTION and a goto function GOTO.

A configuration of an LP parser is a pair whose first component is the stack contents and whose second component is the unexpended input:

\[
(s \times s \times s \ldots \times s, a a \ldots a s)
\]

The next move of the parser is determined by reading a, the current input symbol, and s, the state on top of the stack, and by consulting the parsing action table entry ACTION[s, a]. The configuration resulting after each of the four types of moves are as follows:

1. If ACTION[s, a] = shift s, the parser executes a shift move, entering the configuration

\[
(s \times s \times s \ldots \times s a s, a a \ldots a s)
\]

Here the parser has shifted the current input symbol a and the next state s=GOTO[s, a] onto the stack; a becomes the new current input symbol.
2. If \( \text{ACTION}(s, a, l) = \text{reduce } A \rightarrow \beta \), then the parser executes a reduce move, entering the configuration

\[(s \times s \times s \ldots \times s \times A, a, a, \ldots, a, s)\]

\[0 \ 1 \ 1 \ 2 \ 2 \ m-r \ m-r \ 1 \ 1+1 \ n\]

Where \( s = \text{GOTO}(s, A) \) and \( r \) is the length of \( \beta \). Here the parser first popped \( 2r \) symbols off the stack, exposing state \( s \). The parser then pushes both \( A \), the left side of the production, and \( s \), the entry for \( \text{ACTION}(s, m-r, A) \), onto the stack. The current symbol is not changed in a reduce move.

3. If \( \text{ACTION}(s, a, l) = \text{accept} \), parsing is completed.

4. If \( \text{ACTION}(s, a, l) = \text{error} \), the parser has discovered an error and calls the error procedure.

The parsing is very simple. Initially, the LALR parser is in the configuration \((s, a, a, \ldots, a, s)\) where \( s \)

\[0 \ 1 \ 2 \ n\]

is a designated initial state and \( a, a, \ldots, a \) is the string \( 1 \ 2 \ n \) to be parsed. The parser executes moves until an accept or error action is encountered.
5. IMPLEMENTATION OF THE PARSING ALGORITHM

The implementation of the algorithm is written in PASCAL. Throughout the discussion of the program, we shall use the following LALR(1) grammar G as an example:

1. $S \rightarrow CC$
2. $C \rightarrow cC$
3. $C \rightarrow d$

The program is divided into four parts, input and FIRST set, sets of items, parsing table, and the parsing process. We will discuss these parts in this chapter.

5.1 Input and FIRST set

The data structure employed in this part is as follows:

TSET = SET OF TERMINALS;
NSET = SET OF NONTERMINALS;
RIGHTSTRING = ARRAY [1..RIGHTPART] OF GRRANGE;
PRODUCTION = RECORD
  LEFT: NONTFRMTNALS;
  RIGHT: RIGHTSTRING;
  LN: INTEGER;
END;
GRAMMAR = RECORD
  AR: ARRAY [0..NOPROD] OF PRODUCTION;
  LENGTH: INTEGER;
END;
FIRSTSETS = ARRAY [NONTERMINALS] OF TSET;

There are three important procedures in this section.
5.1.1 Procedure GETPROD

Procedure GETPROD is to read the productions from the input file, it will designate the first symbol as the start symbol S, and it will form the augment grammar for the grammar read by adding a new production S' --> S, i.e., production number 0. The new grammar G' is:

1. S' --> S
2. S --> CC
3. C --> cC
4. C --> d

The nonterminals will be those symbols formed on the left side of the productions and are represented by the set NST. LAMBDA_SET set holds the nonterminals that produce the empty string \( \lambda \). Finally, the productions are stored in an array represented by the variable GR from type GRAMMAR. Procedure GETPROD is as follows:

PROCEDURE GETPROD;

VAR
  N: GRRANGE;
  I, J: INTEGER;
  CH: CHAR;

BEGIN
  I := 0;
  READ(INPUT, CH);
  CHAR_TO_NUM(CH, N);
  GR. AR[I]. LEFT := 70;
  GR. AR[I]. RIGHT[1] := N;
  GR. AR[I]. LN := 1;
WHILE NOT EDF(INPUT) DO
BEGIN
READ(INPUT, CH);
I := I + 1;
J := 0;
GR. AR[I]. LEFT := N;
NST := NST + [N];
WHILE NOT EOLN(INPUT) DO
BEGIN
READ(INPUT, CH);
CHAR_TO_NUM(CH, N);
J := J + 1;
GR. AR[I]. RIGHT[J] := N;
END;
GR. AR[I]. LN := J;
READLN(INPUT);
READ(INPUT, CH);
IF NOT EDF(INPUT) THEN
CHAR_TO_NUM(CH, N);
IF GR. AR[I]. RIGHT[1] = 0 THEN
LAMBDASET := LAMBDASET + fGR. AR[I]. LEFT);
END;
GP. LENGTH := I;
END {GETPROD};

5.1.2 Procedure FIRSTPROD

To find FIRST(A), where A is a nonterminal, we need procedure FIRSTPROD. This procedure uses the properties of the FIRST function to find FIRST(A). If the first symbol of the production's right side is a terminal, then this symbol is in FIRST(A). Otherwise, if it is a nonterminal and it is not in LAMBDASET then FIRST(A) includes FIRST(this symbol), else procedure FIRSTPROD is called again recursively with the same production but ignoring the first symbol in the right side.

Example:
Let $P = A \rightarrow E+T$, $E$ in $\Lambda\mu\sigma\nu\delta\alpha\sigma\iota\tau\eta\zeta$, then after calling $\text{FIRSTPROD}(P)$, $\text{FIRSTPROD}$ is called again recursively with a new parameter $P' = A \rightarrow +T$.

To find the FIRST set for all the nonterminals, we repeat procedure $\text{FIRSTPROD}(A)$ until no more terminals are added to the FIRST set. This is accomplished through procedure $\text{GETFIRST}$. Procedures $\text{FIRSTPROD}$ and $\text{GETFIRST}$ are as follows:

PROCEDURE $\text{FIRSTPROD}(\text{PROD}: \text{PRODUCTION})$;

VAR
   $\text{TEMPPROD}: \text{PRODUCTION}$;
   $\text{LP}: \text{NONTERMINALS}$;
   $I: 1.. \text{RIGHTPART}$;
BEGIN
   $\text{LP} := \text{PROD}. \text{LEFT}$;
   IF $\text{PROD}. \text{RIGHT}[1] \leq \text{TULIMIT}$
      THEN
         $\text{FR}[\text{LP}] := \text{FR}[\text{LP}] + [\text{PROD}. \text{RIGHT}[1]]$
      ELSE
         BEGIN
            IF $\text{PROD}. \text{RIGHT}[1] \in \Lambda\mu\sigma\nu\delta\alpha\sigma\iota\tau\eta\zeta$
               THEN
                  BEGIN
                     IF $\text{PROD}. \text{LN} = 1$
                        THEN
                           $\text{FR}[\text{LP}] := \text{FR}[\text{LP}] + \text{FR}[\text{PROD}. \text{RIGHT}[1]]$
                        ELSE
                           BEGIN
                              $\text{TEMPPROD}. \text{LEFT} := \text{PROD}. \text{LEFT}$;
                              FOR $I := 1 \text{ TO } (\text{PROD}. \text{LN} - 1) \text{ DO}$
                                 $\text{TEMPPROD}. \text{RIGHT}[I] := \text{PROD}. \text{RIGHT}[I + 1]$;
                              $\text{TEMPPROD}. \text{LN} := (\text{PROD}. \text{LN} - 1)$;
                              $\text{FIRSTPROD}(\text{TEMPPROD})$;
                           END
                  END
               END
         END
   END
END

22
END;
END
ELSE
END;
END {FIRSTPROD};

PROCEDURE GETFIRST;

VAR
    TEMP: FIRSTSETS;
    OK: BOOLEAN;
    I, J: 1..NOPROD;
    NI: NONTERMINALS;

BEGIN
    J := GR. LENGTH;
    REPEAT
      OK := TRUE;
      TEMP := FR;
      FOR I := 1 TO J DO
        FIRSTPROD(GR. AR[I]);
      NI := NLLIMIT;
      FOR NI := NLLIMIT TO NULLIMIT DO
        IF TEMP[NI] <> FP[NI] THEN
          OK := FALSE;
      UNTIL OK;
    END {GETFIRST};

5.1.3 Procedure FIND_LASET

In building sets of LALR(1) items, if \([A \rightarrow \lambda, B\beta, a] \) in I, then for each production \(B \rightarrow \gamma\), and terminal \(b\) in FIRST(\(\beta a\)), we add \([B \rightarrow \gamma, b]\) to I. Procedure FIND_LASET finds the lookahead set for the LR(0) item \(B \rightarrow \gamma\). Procedure FIND_LASET is as follows:

PROCEDURE FIND_LASET(S: RIGHTSTRING;
    LAS1: TSET; LIN: INTEGER; VAR LAS2: TSET);

VAR
TEMPS: RIGHTSTRING;
I: 1.. RIGHTPART;
M: INTEGER;

BEGIN
IF S[1] <= TULIMIT THEN
IF S[1] = 0 THEN
LAS2 := LAS2 + LAS1
ELSE
LAS2 := LAS2 + [S[1]]
ELSE BEGIN
IF S[1] IN LAMBDASET THEN
IF LIN = 1 THEN
LAS2 := LAS2 + (FR[S[1]] - [0]) + LAS1
ELSE BEGIN
LAS2 := LAS2 + FR[S[1]] - [0];
M := LIN - 1;
FOR I := 1 TO M DO
TEMPS[I] := S[I + 1];
FIND_LASET(TEMPS, LAS1, M, LAS2);
END
ELSE
LAS2 := LAS2 + FR[S[1]];
END;
END (FIND_LASET);

5.2 Sets of Items

The data structure employed in this section is as follows:

ITEM = RECORD
PRND: 0.. NOPROD;
DOT: 0.. RIGHTPART;
LASET: TSET;
GOO: 0.. NOSTATES;
END;
STATE = RECORD
ARY: ARRAY [1.. MAXITEMS] OF ITEM;
STATELEN: INTEGER;
END;
ITEMSETS = RECORD
    ARR: ARRAY [0.. NONSTATES] OF STATE;
    LEN: INTEGER;
    END;

There are two important procedures in this section:

5.2.1 procedure CLOSURE

In chapter 4, an algorithm for constructing procedure CLOSURE(I) was given. Procedure Closure which we are discussing in this section is not exactly the same. Given the item \([A \longrightarrow \alpha .B \beta , a]\) in \(I\), then procedure CLOSURE only adds the item \([B \longrightarrow .\gamma , b]\) to \(I\) if it is not already there.

For example: In the grammar \(G\) given the item \([C \longrightarrow .c.C, s]\), then procedure CLOSURE adds the items
\([C \longrightarrow .c.C, s]\)
\([C \longrightarrow .d, s]\)
to state \(I\).

Procedure CLOSURE is as follows:

```pascal
PROCEDURE CLOSURE(ITEM: ITEM; VAR ST: STATE);
VAR
    H, I, J, K, L, M: 0.. MAXITEMS;
    LL: INTEGER;
    NN: NONTERMINALS;
    SS: RIGHTSTRING;
    NOTTHERE: BOOLEAN;
```
BEGIN
  J := TEM. DOT + 1;
  K := TEM. PRNO;
  IF GR. AR[K]. LN < (J + 1)
  THEN
    BEGIN
      SS[1] := 0;
      LL := 1;
      END
  ELSE
    BEGIN
      FOR H := (J + 1) TO GR. AR[K]. LN
      DO
        BEGIN
          M := H - J;
          SS[M] := GR. AR[K]. RIGHT[H];
        END;
      END;
      NN := GR. AR[K]. RIGHT[J];
      FOR I := 1 TO GR. LENGTH
      DO
        IF GR. AR[I]. LEFT = NN
        THEN
          BEGIN
            NOTTHERE := TRUE;
            FOR L := 1 TO ITEMCOUNT
            DO
              IF (STT. ARY[L]. PRNO = I) AND (STT. ARY[L]. DOT = 0) OR (GR. AR[I]. RIGHT[1] = 0)
              THEN
                BEGIN
                  FIND_LASET(SS, TEM. LASET, LL, STT. ARY[L]. LASET);
                  NOTTHERE := FALSE;
                END;
            IF NOTTHERE
            THEN
              BEGIN
                ITEMCOUNT := ITEMCOUNT + 1;
                WITH STT. ARY[ITEMCOUNT] DO
                  BEGIN
                    PRNO := I;
                    IF (GR. AR[I]. RIGHT[1] = 0) AND (GR. AR[I]. LN = 1)
                    THEN
                      DOT := 1
                    ELSE
                      DOT := 0;
                  END;
              END;
          END;
        END;
    END;
5.2.2 Procedure BUILD_NEWSTATE

The algorithm for building the sets of \( \text{LALR}(1) \) items states to build the sets of \( \text{LR}(1) \) items, then the states with the same core are merged to form a new state. The resulting sets of items are called the sets of \( \text{LALR}(1) \) items. The number of \( \text{LALR}(1) \) sets of items (states) is the same as the number of \( \text{LR}(0) \) sets of items and generally it is much smaller than the number of sets of \( \text{LR}(1) \) items.

For example, for the following \( \text{LALR}(1) \) grammar

\[
\begin{align*}
E & \rightarrow E+T \\
E & \rightarrow T \\
T & \rightarrow T*F \\
T & \rightarrow F \\
F & \rightarrow (E) \\
F & \rightarrow a
\end{align*}
\]

The number of sets of \( \text{LR}(1) \) items is 22, but that for \( \text{LALR}(1) \) is 12. For a language like ALGOL, the LR table would have several thousand states compared to several hundred in the case of an \( \text{LALR} \) language.

This shows how important the LALR parsers are and how space consuming it is to construct LR parsers. If we
build the sets of LP items and then merge the states with identical cores, this will consume a lot of space. To avoid this, during the process of building a new state, if already there is a state with identical cores, the new state is merged immediately. By this means we save a lot of space in the process of building the LALR(1) sets of items. All this is done in procedure BUILD_NEWSTATE.

Example:

For the grammar G, the sets of LALR(1) items are listed below:

0 : S'-->S ,${}$ 1
   S-->CC ,${}$ 2
   C-->CC ,${c,d}$ 3
   C-->d ,${c,d}$ 4

1 : S'-->S ,${}$

2 : S-->CC ,${}$ 5
   C-->CC ,${}$ 3
   C-->d ,${}$ 4

3 : C-->CC ,${c,d,s}$ 6
   C-->CC ,${c,d,s}$ 3
   C-->d ,${c,d,s}$ 4

4 : C-->d ,${c,d,s}$

5 : S-->CC ,${}$

6 : C-->CC ,${c,d,s}$
5.3 Parsing Table

The data structure employed in this section is as follows:

\[
\begin{align*}
\text{ACTION} &= \ (S, R, A, G); \\
\text{ENTRYPTR} &= \ast. \text{PTR3}; \\
\text{PTR3} &= \text{RECORD} \\
&\quad \text{ACT: ACTION;} \\
&\quad \text{N: INTEGER;} \\
&\quad \text{END;} \\
\text{SYMPTR} &= \ast. \text{PTR2}; \\
\text{PTR2} &= \text{RECORD} \\
&\quad \text{SYM: GRRANGE;} \\
&\quad \text{NEXT: SYMPTR;} \\
&\quad \text{ENTRY: ENTRYPTR;} \\
&\quad \text{END;} \\
\text{STATEPTR} &= \ast. \text{PTR1}; \\
\text{PTR1} &= \text{RECORD} \\
&\quad \text{ST: INTEGER;} \\
&\quad \text{NXT: STATEPTR;} \\
&\quad \text{FIRST: SYMPTR;} \\
&\quad \text{END;}
\end{align*}
\]

The direct way to construct the parsing table is to represent it by a matrix (two dimensional array). By this method it is faster to access the entries in the table, but it is space consuming because most of the table entries are error entries. The other way is to build the table dynamically using pointers as indicated above in the data structure. Here a lot of space is saved, but a longer access time is obtained. Fig. 5.1 shows the LALR(1) parsing table for the grammar G.
During the process of building the table, the function SEARCH tells if there is a conflict in a certain state of the table, which means that the given grammar is not LALR(1). The function SEARCH is as follows:

FUNCTION SEARCH(R: GRRANGE): BOOLEAN;

VAR
  PT: SYMPTR;
  YES: BOOLEAN;
BEGIN
  YES := FALSE;
  PT := SYMBEG;
  WHILE PT^. NEXT <> SYMEND DO
    BEGIN
      PT := PT^. NEXT;
      IF PT^. SYM = R THEN
        BEGIN
          YES := TRUE;
        END;
    END;
END.
5.4 Parsing Process:

The data structure employed in this section is as follows:

TYPE
STACKTYPE = ARRAY [1.. 50] OF INTEGER
;
INPUTSTRING = RECORD
    STRING: ARRAY [1.. 20] OF CHAR;
    LENGTH: INTEGER;
END;

In the program, the parsing process is done interactively. The user feeds the computer terminal with a string to be parsed and waits for the parsing result. The main procedure in this section is procedure PARSING, since the entry ACTION[s,a] can have one of four values:

1. shift
2. reduce
3. accept
4. error

We can summarize procedure PARSING by the following pseudo pascal code:
BEGIN
READ input string;
consult table;
WHILE action <> accept DO
BEGIN
  IF action=shift THEN
    call SHIFT
  ELSE
    call REDUCE;
    consult table;
  END;
END;
print an accepting message;
END;

If an error is encountered, the program stops and prints an error message. Procedure SHIFT shifts the current input symbol on top of the stack and pushes the new state on top of the stack. Procedure SHIFT is as follows:

procedure SHIFT;
BEGIN
  stackptr:=stackptr+1;
  stack[stackptr]:= current symbol;
  stackptr:=stackptr+1;
  stack[stackptr]:= new state;
END;

If the action is to reduce $A \rightarrow \beta$, then procedure REDUCE first pops 2r symbols off the stack (r is the length of $\beta$), then pushes both A, the left side of the production, and s, the new state onto the stack. Procedure REDUCE can be represented as follows:

procedure REDUCE;
BEGIN
  IF $\beta \neq \lambda$ (empty string) THEN
stackptr:=stackptr-2*len of +1
ELSE
  stackptr:=stackptr+1;
  stack[stackptr]:=A;
  stackptr:=stackptr+1;
  stack[stackptr]:= new state;
END;

Fig. 5.2 shows the moves of the LALR(1) parser for the grammar G on the input string cccd:

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>ccd$s shift</td>
</tr>
<tr>
<td>0c3</td>
<td>cdds shift</td>
</tr>
<tr>
<td>0c3c3</td>
<td>dds shift</td>
</tr>
<tr>
<td>0c3c3d4</td>
<td>d$s reduce</td>
</tr>
<tr>
<td>0c3c3C6</td>
<td>d$s reduce</td>
</tr>
<tr>
<td>0c3C6</td>
<td>d$s reduce</td>
</tr>
<tr>
<td>0C2</td>
<td>d$s shift</td>
</tr>
<tr>
<td>0C2d4</td>
<td>$ reduce</td>
</tr>
<tr>
<td>0C2C5</td>
<td>$ reduce</td>
</tr>
<tr>
<td>0S1</td>
<td>$ accent</td>
</tr>
</tbody>
</table>

Fig. 5.2
6. CONCLUSIONS

The parsing algorithm which was analyzed is effective and simple in constructing LALR parsers for this class of context free grammars. In many ways the structure of this algorithm provides more ease and directness in its implementation compared to other LALR parsing algorithm. It is apparent from the algorithm that if a grammar fails to generate the parsing table then it is not LALR(1).

Generally, LR parsers can be constructed to recognize virtually all programming language constructs for which context free grammars can be written.

The complete computer program developed to implement the parsing algorithm was written in the language PASCAL. It is filed with professor Samuel L. Gulden at the division of Computing and Information Science, Lehigh University, Bethlehem, Pennsylvania.
7. LIST OF REFERENCES


8. VITA

The author was born to Mr. and Mrs. Mousa Ali Jaber on October 03, 1955 in Jerusalem, Jordan. He earned his B. Sc. in Mathematics from University of Jordan (Amman, Jordan) in June 1977. In the fall 1979, he began graduate study in Mathematics at Lehigh University and earned his MS. degree in Mathematics in June 1981. In the Fall 1981, he began graduate study in Computing Science at Lehigh University. He was a teaching assistant in Mathematics for the Department of Mathematics at Lehigh University.