AIN - Sprachkonzepte

Oskar Borkenhagen

WS2022/23

Bericht zu den Übungsaufgaben der Vorlesung Sprachkonzepte - AIN

Aufgabe 1

a)

Schreiben Sie ein Java-Programm, das in einem String Formatspezifikationen gemäß *java.util.Formatter* findet.
Erstellen Sie dazu mit der Syntax von *java.util.regex.Pattern* einen regulären Ausdruck für eine solche Formatspezifikation.
Sie brauchen darin nicht zu berücksichtigen, dass bestimmte Angaben innerhalb einer Formatspezifikation nur bei bestimmten Konversionen erlaubt sind. Achten Sie aber bei argment_index, width und precision darauf, ob der Zahlbereich bei 0 oder 1 beginnt.

Beispieleingaben:
xxx %d yyy%n
xxx%1$d yyy
%1$-02.3dyyy
Wochentag: %tA Uhrzeit: %tT

Beispielausgaben:
TEXT("xxx ")FORMAT("%d")TEXT(" yyy")FORMAT("%n")
TEXT("xxx")FORMAT("%1$d")TEXT(" yyy")
FORMAT("%1$-02.3d")TEXT("yyy")
TEXT("Wochentag:")FORMAT("%tA")TEXT("Uhrzeit:")FORMAT("%tT")

a - Lösung

```java
private static String formatter(String input) {
    Pattern patternGeneral =
            Pattern.compile(
                    "(%([1-9]\\$)?[-+#0,(\s]?\\d*(\\.\\d)?[bBhHsScCdoxXeEfgGaA%n])"
            );
    Pattern patternDate =
            Pattern.compile(
                    "(%([1-9]\\$)?[-+#0,(\s]?\\d*[tT][HIklLMSpQZzsBbhAaCYyjmdeRTrDFc])"
            );
    Pattern patternLeftover =
            Pattern.compile(
                    "(%[-+#0,(\s]?\\d*\\D)"
            );
    Pattern usePattern = Pattern.compile(
            patternGeneral.pattern()
                    + "|" + patternDate.pattern()
                    + "|" + patternLeftover.pattern()
    );

    var builder = new StringBuilder();

    Map<String, String> parts = new
        TreeMap<>(Comparator.comparing(input::indexOf));

    Arrays.stream(input.split(usePattern.toString()))
            .forEach(x -> parts.put(x, "TEXT(\"" + x + "\")"));

    usePattern.matcher(input).results()
            .forEach(x -> parts.put(x.group(), "FORMAT(\"" +
                x.group() + "\")"));

    parts.forEach((x, y) -> builder.append(y));

    return builder.toString();
}
```

Pattern realisiert anhand Java Formatter Docs.
Anschließend wird der String in einzelne Teile zerlegt, die dann in einer Map gespeichert werden. Sortiert wird anhand der Position des Keys im Input-String.
Die fertige Map wird dann in einen String umgewandelt.

b)

Erkennen Sie mit ANTLR 4 Lexer-Regeln Zeitangaben im digitalen 12-Sunden-Format gemäß https://en.wikipedia.org/wiki/12-hour_clock. Beachten Sie auch die alternativen Schreibweisen 12 midnight und 12 noon. Testen Sie mit *org.antlr.v4.gui.TestRig*.

b - Lösung

```
1   // TimeLexer.g4
2   lexer grammar TimeLexer;
3
4   Time12H: Default|Noon|Midnight;
5
6   fragment Default:
        ('12:00'|(([1-9]|'1'[01])':'[0-5][0-9]))WS[ap]'.m.';
7   fragment Noon: 'Noon'|'12 noon';
8   fragment Midnight: 'Midnight'|'12 midnight';
9
10  WS: [ \t\r\n]+ -> skip;
```

Lexer Grammatiken beschreiben die Token, die vom Lexer erkannt werden sollen. Fragmente sind Teile der Grammatik, die nicht direkt erkannt werden, sondern nur in anderen Regeln verwendet werden. Der Ansatz hier war die Zeitangaben in drei Teile zu zerlegen:

- Default: Volle Uhrzeitangaben im klassischen 'HH:MM' Format mit AM/PM Angabe
- Noon: Zusätzlich die Mittagszeit '12 noon' und 'Noon'
- Midnight: Synchron dazu Mitternacht '12 midnight' und 'Midnight'

Noon und Midnight sind hierbei die Ausnahme, aber vorgegeben durch die Aufgabenstellung.

Alternativ könnte man mehr Token beschreiben:

```
// TimeLexerV2.g4
lexer grammar TimeLexerV2;

TIME : HOUR SEPERATOR MINUTE (AM | PM)
| TWELVE SEPERATOR '00' (AM | PM)
| TWELVE 'noon'
| TWELVE 'midnight'
| 'Noon'
| 'Midnight';

TWELVE : '12';

HOUR : '1'[0-1]|[0-9];
MINUTE : [0-5][0-9];
SEPERATOR : ':';

AM : 'a.m.' ;
PM : 'p.m.' ;

WS: [ \t\r\n]+ -> skip;
```

Allerdings wird die Lexer Grammatik hier etwas missbraucht, da die 'TIME' Regel eher als Parser Regel genutzt wird. Lexer Regeln sollten eigentlich nur die Token beschreiben, die vom Lexer erkannt werden sollen.

Aufgabe 2

a)

Denken Sie sich eine kleine Sprache aus. Definieren Sie deren Vokabular mit einer ANTLR4 lexer grammar und deren Grammatik mit einer ANTLR4 parser grammar. Erzeugen Sie für einige Beispieltexte mit Hilfe von *org.antlr.v4.gui.TestRig* den Ableitungsbaum (Parse Tree).

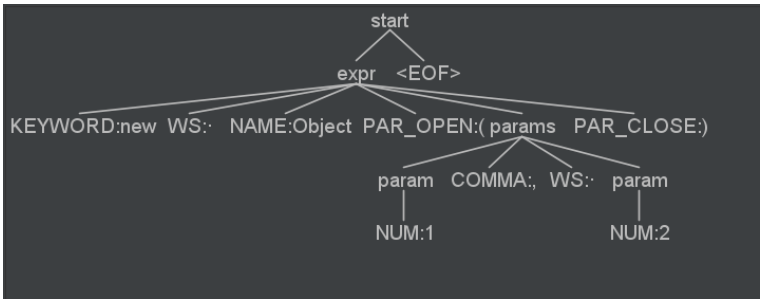
a - Lösung

Dargestellt ist der Lexer einer Sprache, welche die korrekte Kreation einer Java Klasse darstellt. Erlaubt sind Variablen - also einzelne Buchstaben, sowie Zahlen. Parameter werden durch Komma getrennt und dürfen eigene neu erzeugte Klassen sein.

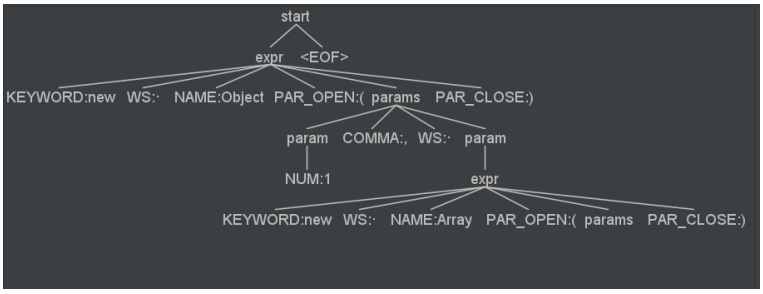
```
1   // CreationLexer.g4
2   lexer grammar CreationLexer;
3
4   KEYWORD : 'new' ;
5
6   NAME : [A-Za-z]+ ;
7   NUM : [0-9]+ ;
8
9   COMMA : ',' ;
10
11  PAR_OPEN : '(' ;
12  PAR_CLOSE : ')' ;
13
14  WS : [ \t\r\n]+ ;
15
16  InvalidChar: . ;
```

Der dazugehörige Parser:

```
1   // CreationParser.g4
2   parser grammar CreationParser;
3   options { tokenVocab=CreationLexer; }
4
5   start : expr EOF;
6
7   expr : KEYWORD WS NAME PAR_OPEN params PAR_CLOSE ;
8
9   params : (param (COMMA WS? param)*)? ;
10
11  param : (expr | NAME | NUM) ;
```



(a) Input: new Object(1, 2)



(b) Input: new Object(1, new Array())

Figure 1: Parse Tree Beispiele

Der Parser ist für die grammatikalische Anordnung der durch den Lexer vorgegebenen Token verantwortlich.

b)

Definieren Sie mit Java-Klassen die abstrakte Syntax Ihrer Sprache aus a) und schreiben Sie ein Java-Programm, das den ANTLR4 Parse Tree in einen AST überführt.

b - Lösung

Definition als UML-Diagramm:

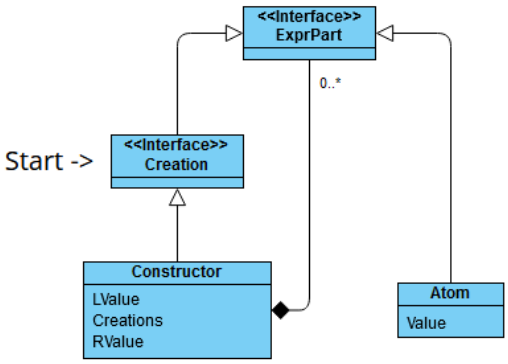


Figure 2: UML-Diagramm

Eine Baumstufe kann in drei Teile geteilt werden:

- **Links**: Der Keyword-, Namen- und Klammer-Teil vor der Mitte.
- **Mitte**: Rekursive Parameter-Definition - kann auch leer oder eine neue *Creation* sein.
- **Rechts**: In diesem Fall die Klammer zum Schließen der Parameter.

Dadurch ergibt sich die Vorgabe von mindestens einer Constructor Instanz. *L-* und *RValue* sind definierbare Strings und *Creations* ist die Liste der Parameter innerhalb der Klammern/Strings, dargestellt in der Relation. So können neue *Constructor* Objekte, sowie Werte definiert in der Klasse *Atom* als Parameter übergeben werden.

Um die Struktur beizubehalten, werden beispielhaft also folgende Klassen erstellt:

```
1  public interface Expr {
2  }
3
4  public interface Creation extends Expr {
5  }
6
7  public class Atom implements Expr {
8      private final String val;
9
10     public Atom(String val) {
11         this.val = val;
12     }
13     public String getVal() {
14         return val;
15     }
16     @Override
17     public String toString() {
18         return this.val;
19     }
20 }
21
22 public class Constructor implements Creation {
23     private final String leftVal;
24     private final List<Expr> params;
25     private final String rightVal;
26
27     public Constructor(String leftVal, List<Expr> params, String
           rightVal) {
28         this.leftVal = leftVal;
29         this.params = params;
30         this.rightVal = rightVal;
31     }
32     public String getLeftVal() {
33         return leftVal;
34     }
35     public List<Expr> getParams() {
36         return params;
37     }
38     public String getRightVal() {
39         return rightVal;
40     }
41     @Override
```

```
42      public String toString() {
43          return this.leftVal + this.params + this.rightVal;
44      }
45  }
```

Um diese Struktur in einem Builder umzusetzen, muss ein eigener Stack für jedes neue Constructor Objekt angelegt werden.

```
1   public class CreationBuilder extends CreationParserBaseListener {
2
3       private final List<Stack<Expr>> stackList = new LinkedList<>();
4       private int depth = -1;
5
6       public Creation build(ParseTree tree) {
7           new ParseTreeWalker().walk(this, tree);
8
9           return (Creation) this.stackList.get(this.depth).pop();
10      }
11
12      @Override
13      public void enterExpr(CreationParser.ExprContext ctx) {
14          this.stackList.add(new Stack<>());
15          this.depth++;
16      }
17
18      @Override
19      public void exitExpr(CreationParser.ExprContext ctx) {
20          if (ctx.getChildCount() == 6) {
21              var l = new StringBuilder();
22              for (int i = 0; i < 4; i++) {
23                  l.append(ctx.getChild(i).getText());
24              }
25
26
27              var c = new Constructor(
28                      l.toString(),
29                      new LinkedList<>(this.stackList.get(this.depth)),
30                      ctx.getChild(5).getText()
31              );
32              this.stackList.get(this.depth).clear();
33
34              if (this.depth > 0)
35                  this.depth--;
36              this.stackList.get(this.depth).push(c);
37
```

```
38          }
39      }
40
41      @Override
42      public void enterParam(CreationParser.ParamContext ctx) {
43          if (ctx.start.getType() == CreationParser.NUM) {
44
45              this.stackList.get(this.depth).push(new
                    Atom(ctx.NUM().getText()));
46
47          } else if (ctx.start.getType() == CreationParser.NAME) {
48
49              this.stackList.get(this.depth).push(new
                    Atom(ctx.NAME().getText()));
50
51          }
52      }
53
54  }
```

Für jedes angefangene Constructor Objekt wird ein neuer Stack angelegt in der *enterExpr* Methode. Zusätzlich wird die Tiefe erhöht, um jeweils beim aktuell noch nicht abgeschlossenen Constructor Objekt zu bleiben. Dadurch bleibt die Reihenfolge erhalten und innerhalb eines noch nicht abgeschlossenen Constructor Objekts können *Atom* Objekte und weitere *Constructor* Objekte erstellt werden, ohne dass Konstruktoren die auf dem Stack liegenden Objekte in sich aufnehmen.

Die *exitExpr* Methode arbeitet den aktuellen Stack ab und setzt ihn als Parameter-Liste für das aktuelle *Constructor* Objekt. Danach wird der aktuelle Stack geleert und die Tiefe verringert. Abschließend wird der aktuelle Konstruktor auf den Stack gelegt für den nächsten Konstruktor oder den finalen Rückgabewert.

Die *enterParam* Methode ist der unterste Baustein des Baums, und dient zur Erstellung von *Atom* Objekten, die dann auf den Stack gelegt werden für Konstruktoren. Wenn die *Param* Regel wieder eine *Expr* Regel enthält, wird sie in der Methode ignoriert, da diese bereits in der *exitExpr* Methode abgearbeitet wird.