Montages:
Modular Specifications of Programming Languages

Philipp W. Kutter
Eidgenössische Technische Hochschule
CH–8092 Zürich (Switzerland)
kutter@tik.ee.ethz.ch

Alfonso Pierantonio
Università di L'Aquila
I–67100 L'Aquila (Italy)
alfonso@univaq.it

Abstract

Montages are a new way of describing all aspects of programming languages formally. Such specifications are intelligible for a broad range of people involved in programming language design and use. In order to enhance readability we combine visual and textual elements to yield specifications similar in structure, length, and complexity to those in common language manuals, but with a formal semantics. The formal semantics is based on the well known Gurevich's Abstract State Machines (formerly called Evolving Algebras).

1 Introduction

In recent years, programming languages followed the trend towards higher levels of abstraction. As a consequence, their design became more and more a non trivial team activity. They could often not retain compactness and coherence of languages like Pascal which are designed by a single person. Thus, language design appears to be on the verge of being considered a systematic engineering process. New languages are defined through a number of stages from initial design to routine–use by programmers, the so called *programming language life cycle*.

As for other software artifacts, accurate, consistent and intellectually manageable descriptions of languages are needed for different reasons. For instance, designers need to keep track of already taken decisions about language constructs, and the design intentions must be conveyed to the implementors. So far, the only comprehensive description of a programming language is likely its reference manual, which is mainly informal and open to misinterpretation. Formal approaches are therefore sought. Nevertheless, they are regarded very warily. Often they caused a certain dissatisfaction because of the pragmatic side of language specification.

Unlike other formalisms which base both static and dynamic semantics directly on syntax, Montages base both of them on a control and data flow analysis, which is specified as well. The semantics of all specification components is formally defined using Gurevich's Abstract State Machines (GMs). GMs have been proposed by Y.Gurevich as a dynamic generalization of multi–sorted algebras, intended to provide a more versatile notion of Turing machine, "able to simulate arbitrary algorithms in a direct and essentially coding-free way" [Gur95]. In short, GMs are a state–based formalism in which a state is updated in discrete time steps. Unlike most state based systems, the state is given by an algebra, that is, a collection of functions and universes. The state transitions are given by rules that update functions pointwise and extend universes with new elements.

GMs have already been successfully used to model the dynamic semantics of programming languages and systems such as C [GH93], C++ [Wal94], Oberon [Kut97], Occam [BD96], Prolog [BR95], and VHDL [BGM95]. General aspects based on such models have been investigated in [GR94, GKS90]. At the risk of oversimplifying somewhat, we can describe some of these models [GH93, Wal94, Kut97] as follows. Program execution is modeled by the evolution of two functions CT and S. CT points to the part of the program text currently in execution and may be seen as an abstract program counter. S represents the current value of the store. Formally one defines the *initial state* of the functions and specifies how they evolve by means of *transition rules*.

In the mentioned models [GH93, Wal94, Kut97], the *initial state* is assumed to include a representation of the program's control and data flow in the form of functions between parts of the program text. As usual the control flow functions specify the order in which statements are executed, and the data flow functions specify how values flow via variables through operations. The corresponding *transition rule* of the models updates the program counter and program state using the control and data flow functions.

The definition of the functions representing control and data flow information is the result of a static analysis, which is only described informally. The lack of formalization of the static analysis remains an heavy limitation to the completeness of the given specification. Moreover, it prevents the specification from being executed using existing GM–interpreters [DCIG93].

Nevertheless, GMs provide compact and natural models of program execution for several reasons:

- By using a state–based framework, they account directly for the use of state in imperative languages.
- By representing control– and data–flow explicitly, they provide a more understandable and compact description of the dynamics than traditional operational and denotational methods, which operate only on abstract syntax trees.
- They allow us to model the language at an appropriate level of abstraction. For instance, the store may contain abstract objects (such as arrays or records) that directly reflect the type system of the language.
- Parallelism is easily covered by generalizing the abstract program counter to a set of active tasks.

The contribution of Montages is intended twofold. On one hand, we intend to complete the classical GM approach accounting for the static analysis as well, i.e. describing in an intelligible yet formal and abstract way how to define the control and data flow starting from the parse tree. This is mainly reached through the use of a formal graphical language. On the other hand, we aim at a formalism which can be equally well understood by language designers, compiler constructors and programmers. Each language construct is specified separately and the overall complexity is decomposed according to the syntax. Such a syntax–driven decomposition of the overall complexity provides for a very neat localization of the behavior. Interdependencies are drawn formally enforcing the independence of the design of each single construct. The resulting system of loosely coupled and highly cohesive components facilitates language development and improves language understanding and maintenance.

The reader at this point may be keen to know whether realistic languages can be specified adequately with such an approach. Due to space limitations this is hard to show in this paper but we invite the reader to take a look at relevant full–scale case studies. For instance in [KP97], which is available on the net, we gave the Montage specification of the Oberon programming language. Such an experience suggests that messy features like pointers, gotos, type checking and name analysis in complex situation are surprisingly short, comprehensive and elegant. A second complex case study is the specification of SQL [DiF97]. This work shows how not only imperative languages may be described. Finally, Montages are used at ICSI, Berkeley for the specification of the object–oriented language Sather [Anl97]. Sather is a typical example of a languages designed and maintained by a team. One of the goals of the specification of Sather with Montages is to compare two different dialects of the languages, trying to understand where the semantics differs apart from syntax differences.

The paper is organized as follows. In section 2, we give prerequisites needed for the formal semantics of Montages which is given in section 3. In section 4, we show some advanced features of Montages, which are used in the large case studies. Finally, in the last two sections, we provide a brief comparison with related work and draw conclusions.

2 Prerequisites

In the following sections we give some preliminary notions. In section 2.1 we recall the notion of context free grammars, look at the related derivation trees, and introduce *compact derivation trees*. The main point of the section is the observation, that the nodes of compact derivation trees are characterized by a specific subset of the symbols in the grammar. This subset, the so–called *characteristic symbols* is the base for the syntax driven modularity of Montages.

In section 2.2 we introduce the notion of *initial state* of GMs, and specify how a compact derivation tree is mapped in such a state. Finally in section 2.3 we introduce the notion of *transition rule* of GMs, and show some generic transition rules which can be used to traverse a compact derivation tree parallelly and sequentially.

Although all used aspects of GMs are explained during the sections, we have to be rather short. For a more complete treatment and motivations we refer to [Gur95].

2.1 Grammars and Trees

We recall some basic notions (more details can be found in [ASU85]). A context-free grammar G is defined by a finite vocabulary V_t of *terminals*, another finite vocabulary V_n of *nonterminals* with $V_t \cap V_n = \phi$, a *start symbol* $S \in V_n$ that starts all derivations, and a finite set P of *production rules* of the form $n ::= E$ where $n \in V_n$ and E is a regular expression as follows:

- s denotes a string s;

- $[E_1]$ denotes the strings denoted by E_1 and the empty string;
- $E_1|E_2$ denotes the union of the strings denoted by E_1 and E_2;
- E_1E_2 denotes all concatenations of a string denoted by E_1 followed by a string denoted by E_2;
- $\{E_1\}$ specifies all strings obtained by concatenation of zero or more strings denoted by E_1,

where E_1 and E_2 are arbitrary regular expressions over $V_t \cup V_n$.

The language defined by G is the set of strings consisting of only terminals and generated from the start symbol by applying the production rules.

The generation of a string S can be described by means of a *derivation tree*. The root of the tree is labeled with the start symbol. We say as well, the root represents the start symbol. Every replacement of a nonterminal n by $s_1\ s_2\ \ldots\ s_m$ in the derivation of S is represented in the tree by appending from left to right nodes representing s_1 to s_m to the node representing n. The new nodes are labeled with the corresponding symbols s_1 to s_m. Such trees can be made more compact by putting multiple labels in the case of *synonym productions* [Ode89]. Synonym productions are rules of the form $n ::= s_1|s_2|\ldots|s_m$, which give place to nodes with only one child. In such cases we simply do not append new nodes but we keep track of the synonym productions by adding a new label to the current node. The resulting trees are called *compact derivation trees* and we distinguish a synonym production $n ::= E$ by writing $n = E$.

As example for a grammar of a very simple language take the following:

Program	::=	VAR DeclSeq BEGIN StmentSeq END
DeclSeq	::=	DeclIdent {“,” DeclIdent}
DeclIdent	=	ident
Desig	=	DesigIdent | NextAccess | DataAccess
DesigIdent	=	ident
NextAccess	::=	Desig“.”NEXT
DataAccess	::=	Desig“.”DATA
Expr	=	Factor | Sum
Factor	=	intnumber | Desig
...		
StmentSeq	::=	Statement {“;” Statement}
Statement	=	Assignment | NewStment | WhileStment
Assignment	::=	Desig “:=” Expr
NewStment	::=	NEW“(”Desig“)”
...		

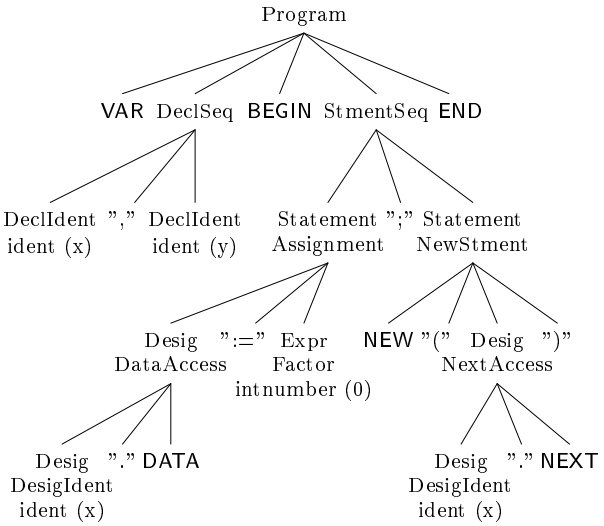


Figure 1: A compact derivation tree

The language features pointer manipulation and a dynamic store which contains only objects of a fixed type. This objects have a NEXT field which can point to other objects and a DATA field which can hold an integer. The variables in the declaration sequence are such objects, and new objects can be allocated with the NEW-statement. Despite of its compactness, this languages can be used for all sorts of linked-list algorithms. The complete semantics is described in [KP96]. The compact derivation tree associated to the following program of the language is in fig. 1

```
VAR x,y
BEGIN
  x.DATA:= 0;
  NEW(x.NEXT)
END
```

According to the above definitions, each node is labeled with at least

- one terminal or
- one non-terminal, which is the left-hand-side of a non-synonym production.

Such symbols are called *characteristic symbols* since it can be shown that each node is labeled with exactly one of them. If a node is labeled with a characteristic symbol s we say as well that the node is characterized by s. Such a characterization partitions the set of nodes.

2.2 Initial State and Tree Representation

Given a program, its compact derivation tree is represented in the associated initial state. A state of a GM is

given by a set called the *superuniverse* and a collection of functions. The superuniverse has the distinguished elements *true*, *false*, and *undef*. Unary functions from the superuniverse to $\{true, false\}$ are used to represent sets or *universes*. The universe consisting of *true* and $false$ is called *Bool*. Our setting requires some specific universes. In particular, the nodes of the compact derivation tree constitute the universe *Node*. Moreover, each symbol s in $V_t \cup V_n$ is interpreted by a sub-universe of *Node* containing those nodes which are labeled by s.

Example The tree in fig. 1 states that universe *Desig* contains four elements, namely all nodes with the label *Desig*.

As mentioned in the last section, the universes interpreting the characteristic symbols partition the universe *Node*, i.e. each node is in exactly one such universe. But a node might be present in more than one universe if it has multiple labels, which is possible if its derivation includes synonym productions.

Example In fig. 1 the left– and bottom–most node is member of the universes *Desig*, *DesigIdent*, and *ident*, where *ident* is its unique characterization.

A number of so–called *selector functions* reflects the structure of compact derivation trees and allows us to retrieve the syntactical elements of the program text. Since descendants of a node are constructed by a production rule, we define the functions accordingly. Let x be a node whose descendants have been constructed by a replacement due to a production rule $n ::= E$, then

- If E is of the form "$s_1 s_2 ... s_m$" we access the new nodes in the universes $s_1, s_2 \cdots$, and s_m by unary functions

  $$(\text{S-}s_i : Node \to s_i)_{i \in \{1,\ldots,m\}}$$

  If the same symbol s occurs more than once in "$s_1 s_2 \cdots s_m$", we enumerate the functions from left to right: S1-s maps x to the first s-descendant, S2-s to the second and so on.

- If E contains a symbols s in a { } part, then an element of a universe *ListNode* is created and serves as access point of the whole list. The details are given in section 4.1.

Apart from the selector functions, which reflect the structure of the tree, we need also an auxiliary function $Up : Node \to Node$, which links the descendants to their parents.

The discussion on the automatic definition of initial state is deferred elsewhere since it is out of the scope of the paper. Nevertheless it is worth mentioning that this task can be again accomplished by an opportune GM which would parse the program and build the tree. Of course, one may alternatively make use of usual tools, such as YACC.

2.3 Transition Rules and Tree Traversal

Given an initial state a transition rule governs the behavior of a GM. A transition rule is a description of how the current state evolves starting from the initial one. The transition rules are built over the following set of constructs: *update*, *block*, *conditional*, *vary* and *extend*. The update operator is a pointwise modification of a function. Its synopsis is

$$f(t_1, \ldots, f_n) := t_0$$

Such a rule first evaluates the terms $t_1, \ldots, t_n$ and t_0 over the current state to the elements $e_1, \ldots, e_n$ and e_0, and then modifies f on the point $(e_1, \ldots, e_n)$ to e_0. The update of f does not affect the definition of the function over the rest of the domain. A set of transition rules constitute a block and are executed in parallel. For convenience we speak henceforth about "transition rules", if we mean the block of them. A conditional rule consists of a condition C and another rule R.

if C **then**
  R
endif

The *then* branch R is triggered if the value of the guard expression C is evaluated to *true*. A conditional rule may have also an *else* branch, in such a case the corresponding transition rule is triggered if the guard is not evaluated to *true*. The operators *vary*, and *extend* are explained as they are encountered in the paper. The different states of an GM are reached by iteratively triggering the transition rule until the GM reaches a state which cannot evolve anymore.

Montages model the static semantics and analysis with a GM that traverses the tree. In the next section we define what exactly happens during this traversal. In this section we define a general pattern of transition rules which can be used to define a traversal. In instance of the pattern traverses a compact derivation tree of a given grammar, and executes at each node an action, which depends on the characterization. In addition a general technique is introduced allowing to sequentialize the tree traversal in an arbitrary way.

Imperative versus declarative style In [KP96] we defined tree traversal in an imperative style. Here we use alternatively a definition in a declarative style. The declarative style presented here can be used for parallel traversal as well, which is needed for the derivation of

parallel compilers. The imperative version of [KP96] is a sequential refinement of the declarative version given here. The advantage of the imperative version is that it is easier to read for non-academic programmers. Another advantage of the version given here, is that it facilitates extraction of an axiomatic description of the state reached after the graph traversal.

As noted the aim of the traversal is to execute for each node in the tree an action. We start with a rule executing the action for all nodes in parallel. Then we show how to specify an action depending on the characterization of the node. Finally we present a solution, how the execution of the actions can be sequentialized.

Parallel traversal In order to execute an action R for each node, we use the vary construct of GMs. The rule

vary CN **over** Node (1)
  R
endvary

executes R for each element in Node simultaneously. The bound variable CN (current node) can be used in R to access the single elements. If for instance Node is a universe with two elements a and b, the above rule corresponds to a block of twice R, once with CN substituted by a and once with CN substituted by b.

Case distinction by characterization Using the fact that the nodes are partitioned by their characterization, we can execute a specialized rule R_n, the so called *action of n*, for a node characterized by n. Technically we can do this by replacing R in the above vary rule by a block of conditionals, one for each characterizing symbol n:

if n(CN) **then** (2)
  R_n
endif

Such a conditional triggers R_n only, if n(CN) evaluates to *true*, i.e. if CN is in the universe n. For convenience we say henceforth action of a node, if we mean the action of its characterization.

Up to now, for each node, its action is executed in parallel. The next task is to introduce the possibility to sequentialize the execution. Typically the actions should be executed for lower level nodes first, and in some order between the children of a node. The situation where actions of lower level nodes are executed first allows already for direct representation of structural induction: each node (representing a parsed term) can use the results of the actions (definitions) performed for the descendants (representing sub-terms). In addition we need often a certain sequentialisation between descendants, e.g. actions for declaration parts in programs must typically be performed before actions for the statement parts.

For the sequentialisation task, we need a boolean dynamic field

$$\textit{Visited: Node} \rightarrow \textit{Bool}$$

which is initialized with false for each node. This field indicates whether a node has been visited, i.e. whether its action has been executed. A relation

$$\textit{before: Node} \times \textit{Node} \rightarrow \textit{Bool}$$

relates nodes sequentially. The relation (*a before b*) indicates node a must be visited before node b. The relation *before* can be defined with a parallel tree traversal, as we will see in the next section.

Sequentialized traversal Using the above definitions, a sequentialized traversal is defined by the following rule

vary CN **over** Node (3)
  satisfying
    for all node **in** Node **holds**
      node before CN **implies** node.Visited
  R
  CN.Visited := true
endvary

where again R is refined to a case distinction by characterization (2). The final state or termination of a sequentialized tree traversal is reached if the root of the tree is visited.

3 Montages

Montages is a semi–visual formalism which defines for each characteristic symbol the related syntactic and semantic aspects of a language. We start by giving two examples for Montages of programming language constructs.

The first example, given in fig. 2, is a typical While loop. The topmost part is the production rule defining the context free syntax. Below is a graphical representation of the control and data flow graph. The NT (NextTask), and TrueTask arrows denote for instance sequential control flow, while the Condition arrow denotes the data flow. Control flow arrows are dotted and data flow arrows are solid. The control flow arrows I (initial) and T (terminal) are special arrows which serve to plug together the local flow–information to the global one. The

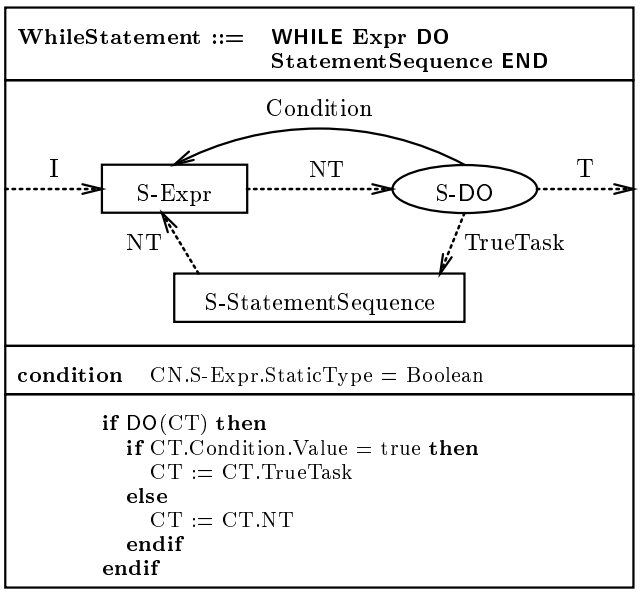


Figure 2: A Montage for a While statement

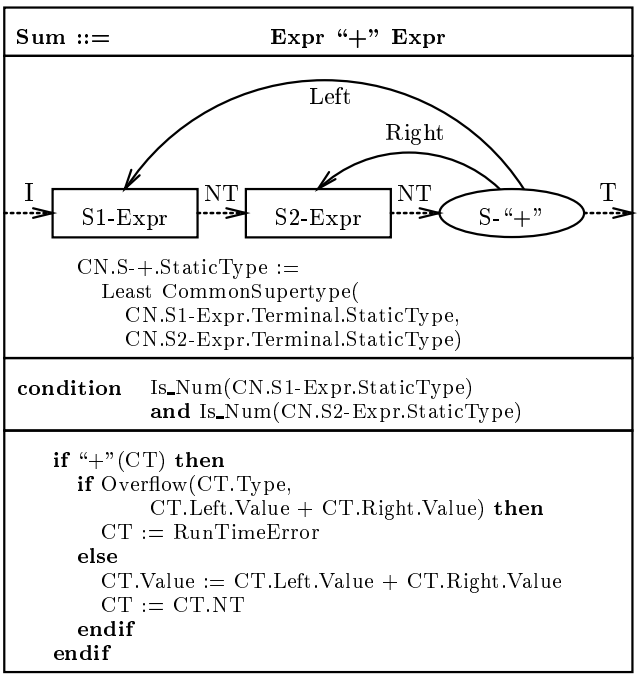


Figure 3: A Montage for a Sum expression

boxes and circles are labeled with the selector functions accessing the corresponding elements. A circle is used to indicate that the corresponding descendant is a token of the program text. The third part of the While Montage contains the static semantics, that is, the type of the While-condition must be Boolean. The last part contains the dynamic semantics rules. This rule is executed if the abstract program counter CT points to a DO-task. In this case, it checks whether the value of the condition is true. If it is true, the abstract program counter is set to the statement sequence (using the TrueTask arrow), else to the next task. The next task of the DO token is not defined directly by the graph, but it is defined through the mentioned plugging mechanism of the T arrow.

As a second example, we show a Sum Montage (fig.3). The static semantics of the sum expresses that both components must be of numeric type. In the definition we use a static function Is_Num(·) which maps all numeric types to true. The graph specifying control and data flow defines again NT control flow arrows, and two data flow arrows Left, Right, which are used to reference the left respectively right argument of the "+" token. In contrast to the first example, the second part of the Sum Montages contains a textual rule. This rule uses the static function LeastCommonSupertype in order to determine the type of the Sum from the types of the left and right arguments. The dynamic semantics rule of the "+"-token raises a runtime error, if the addition leads to an overflow, with respect to the type of the expression. Otherwise the value of the "+"-token is set to the result of the addition and control is passed to the next task.

It is remarkable how the understanding of a Montage does not require too much expertise as shown in the examples above. The formal semantics given below is an unambiguous arbiter between different ways of understanding and it makes clear how the interaction between the Montages works, e.g. how the I and T arrows plug the local flow information together to global control and data flow graphs. The resulting GM semantics is exactly compact as the visual Montages, i.e. each element in a Montages corresponds to an update in the GM-semantics.

Formally speaking, the semantics of a Montages specification is a GM $\mathcal{M}$ that for a given program checks the static semantics, initializes the control and data flow functions, and in a second phase executes the dynamic semantics. The transition rule of $\mathcal{M}$ consists thus of two rules, one modeling the first phase, called *statics rule*, and one modeling the second phase, called *dynamics rule*.

The statics rule is a sequentialized traversal as described in the last section. The actions executed by the statics rule check for each node the static semantics, and define control and data flow between leaf descendants. In the Montages framework, the defined flow information is stored as functions between parts of the program text, i.e. leaves of the derivation tree. This functions

correspond to the arrows in the examples. As usual the final state of such a sequential traversal is reached if the root is visited. If the traversal is aborted, e.g. a function Abort is set to true, the checked program is not valid. Otherwise we can proceed with the dynamics rule.

The second phase uses the final state of the first as its initial state. The dynamics rule executed in the second phase corresponds to the transition rule of a traditional GM for dynamic semantics, as described in the introduction. But in contrast to the traditional use of GMs, where control and data flow are assumed to be given as functions between leaves, here these functions are defined by the first phase.

As exemplified above, Montages are modules containing for a characteristic symbol:

- A production rule, if the symbol is a non-terminal.
- A semi–visual specification of control and data flow functions. The graphical part consists of nodes representing the right-hand-side symbols of the production rule and arcs specifying flow function. The textual part is a transition rule.
- A first–order logic predicate that represents the static semantics constraint. This predicate is marked by the keyword **condition**.
- Transition rules that model the dynamic semantics of terminals generated by the production rule.

These four parts are used to define the statics and dynamics rules. The dynamics rule is simply the block of the rules given in the fourth part of the Montages. The statics rule is a sequentialized traversal (3). The second and third part of a Montage of a symbol s define the action of s in this traversal. The definition of the *before*-relation for the statics rule is done by a parallel traversal (1) with case distinction by characterization (2). The actions of this traversal are defined such that lower nodes in the tree must be visited before higher nodes, and that siblings are visited in the order corresponding to their left-before-right and top-before-bottom order in the graph of the Montages.

We can thus define the s-action in the parallel traversal defining the *before*-relation of the statics rule as follows:

before(CN.S_1, CN.S_2) := true
before(CN.S_2, CN.S_3) := true
...
before(CN.S_{n-1}, CN.S_n) := true
before(CN.S_n, CN) := true

where S_1, S_2, ..., S_n are the selector functions accessing the descendants of an s-node in the left-before-right and top-before-bottom order defined by the Montage of s. Please note that the transitive closure of the defined relation is not necessary due to the way how the sequentialisation mechanism works.

To illustrate we give the corresponding actions for the While and Sum Montages.

Example The action for the While Montage (fig. 2) is

before(CN.S-Expr, CN) := true
before(CN.S-DO, CN) := true
before(CN.S-StatementSequence, CN) := true
before(CN.S-Expr, CN.S-DO) := true
before(CN.S-DO, CN.S-Expr) := true

and the action for the Sum Montage (fig. 3) is

before(CN.S1-Expr, CN) := true
before(CN.S2-Expr, CN) := true
before(CN.S-"+", CN) := true
before(CN.S1-Expr, CN.S2-Expr) := true
before(CN.S2-Expr, CN.S-"+") := true

The actions of the statics rule are explained step by step in the following. Lets assume for the discussion a fixed Montage for a symbol s. The action for this Montage is built up as block of updates. Each arrow in the control and data flow graph defines one update in the action. This update links not directly the graphically related nodes but two of their leaf-descendants. The definition which leaf descendants are linked relies heavily on the definition of two functions

$$\mathit{Initial} : \mathit{Node} \to \mathit{Node} \qquad \mathit{Terminal} : \mathit{Node} \to \mathit{Node}$$

which conceptually denote the first and the last leaf in the control flow between the leaf-descendants of a node. These functions are initialized with the identity on nodes, in order to be well defined for leaves, which serve as their own initial and terminal leaf. The definition on inner nodes is built up inductively during the tree traversal. A dotted arrow, labeled with I (respectively T) denotes the direct descendant, whose definition of Initial (respectively Terminal) has to be copied. In the While montage (fig. 2), for instance, the initial leaf of a WhileStatement-node is the initial leaf of the S-Expr-descendant, and the terminal leaf is the terminal leaf of the S-DO-descendant.

The arrows in the graph define three different kind of updates in the action, one for the above described Initial and Terminal functions, one for data flow arrows,and one for control flow arrows:

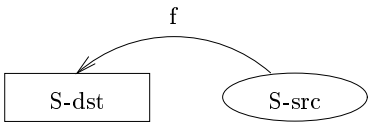
1. To define the functions *Initial* and *Terminal*, we specify which node in the graph contains the initial and terminal leaf, respectively. We call this nodes Inode and Tnode. Inode is defined as the target

of a dotted arrow labeled with I, and Tnode is the source of a dotted arrow labeled with T. Formally these arrows define two selector functions I and T which link the CN with Inode and Tnode. The corresponding fragment of transition rule obtained by specifying Inode and Tnode graphically is the following block:

CN.Initial := CN.I.Initial
CN.Terminal := CN.T.Terminal

We call this *fragment–1.*

2. Each solid edge as the following

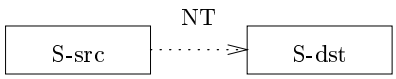


defines the update

CN.S-src.f := CN.S-dst.Terminal

The terminal leaf is chosen as target for the control flow. We call the block of these updates *fragment–2.*

3. Each dotted edge as the following



defines the update

CN.S-src.Terminal.NT := CN.S-dst.Initial

The defined control functions link the terminal leaf with the initial. We call the block of these updates *fragment–3.*

The action of the statics rules contains in addition to the updates corresponding to the arrows a rule which is given textually in the second part of the Montage. Please have a look at the textual rule in the sum Montage (fig. 3). Its structure resembles to that of updates generated by the second fragment, but it cannot be represented graphically. Theoretically all graphical defined updates can be given textually as well.

The only missing part is the check of the static semantics condition. This condition is checked before the updates of the action happen, and a nullary function *Abort* is set to true if the condition is false. In order to make the rule easier to read, we write the corresponding conditional rule at the beginning of all updates. We discussed now all parts of the action of a characteristic symbol, and can define it as follows:

The action of a characteristic symbol s in the sequentialized traversal being the statics rule of the Montages semantics is

if not Condition **then**
  Abort := true
endif
fragment–1.
fragment–2.
fragment–3.
TransRule

where Condition is the static semantics constraint of Montage s, TransRule is the textual rule in the second part of Montage s, and the *fragment–1.*, *fragment–2.*, and *fragment–3.* are the updates defined by the graph of Montage s.

Example We give for the While (fig. 2) and Sum (fig. 3) Montages the corresponding actions. The action for While is the following rule:

if not CN.S-Expr.StaticType = Boolean **then**
  Abort := true
endif

CN.Initial := CN.S-Expr.Initial CN.Terminal := CN.S-DO.Terminal	1
CN.S-DO.Condition := CN.S-Expr.Terminal	2
CN.S-Expr.Terminal.NT := CN.S-DO.Initial CN.S-DO.Terminal.TrueTask := CN.S-StatementSequence.Initial CN.S-StatementSequence.Terminal.NT := CN.S-Expr.Initial	3

and the action for the Sum looks as follows:

if not (Is_Num(CN.S1-Expr.StaticType) **and**
  Is_Num(CN.S2-Expr.StaticType)) **then**
    Abort := true
endif

CN.Initial := CN.S1-Expr.Initial CN.Terminal := CN.S-"+".Terminal	1
CN.S-"+".Left := CN.S1-Expr.Terminal CN.S-"+".Right := CN.S2-Expr.Terminal	2
CN.S1-Expr.Terminal.NT := CN.S2-Expr.Initial CN.S2-Expr.Terminal.NT := CN.S2-"+".Initial	3

```
CN.StaticType :=
    LeastCommonSupertype(
        CN.S1-Expr.StaticType,
            CN.S1-Expr.StaticType)
```

4 Advanced Features

This section is concerned with several aspects related to scaling of Montages to realistic programming languages. First we presents an extension to the visual formalism, which is used for list processing. This extension turned out to be very useful in the case studies. Then we show how parallel and non-deterministic evaluation can be modeled with Montages. Finally the problem of systematic reuse of Montages is tackled by relating the structure of Montages to the concept of object-orientedness.

4.1 List Processing

In many approaches a major part of a language specification is concerned with the processing of lists. Therefore we decided to include in Montages a simple, yet powerful list model together with graphical and textual specification elements that can be used to avoid all explicit list processing.

If the right–hand side of a production rule contains a symbol in a { } part, a list of descendents is generated. As already mentioned we generate as well an additional node, a so called *list node*, that provides access to the elements and to all needed informations about the list. An attribute *ListLength* of the list node is set to the length of the generated list and a binary mix–fix function

$$_[_] : \mathit{ListNode} \times \mathit{Nat} \rightarrow \mathit{Node}$$

can be used to retrieve the elements of the list. The initial and terminal leaves of a list node are defined to be the initial leaf of the first element, respectively the terminal leaf of the last element in the list. If the list is empty, they point to the list node itself, which then serves as dummy element. The dynamic semantics of that dummy element corresponds to the skip command.

For convenience we assume that a number of patterns in the right-hand-side of production rules are recognized and treated as simple lists. These patterns are

{s} s{s} s {“t” s} [s {“t” s}]

where “t” is an arbitrary terminal and *s* a non-terminal. For all these patterns just one list node is generated, which can be accessed by the selector function S-s. The _[_] function can then be used to access all generated *s*-descendants from left to right, regardless by which *s* in the pattern the descendant was generated.

We have not yet defined the statics for list nodes. For simplicity we give just one of many possible solutions. We define the action in the *before*-definition such that the elements must be visited from left to right and that they are visited before the list node CN:

```
vary i over Nat
    satisfying i < CN.ListLength
        before(CN[i], CN[i + 1])
endvary
before(CN[CN.ListLength], CN) := true
```

In addition we chose the action in the statics rule of a list node such, that it always links the elements with a NT control arrows and sets the initial and terminal leaf to the corresponding leaves of the first and last element:

```
CN.Initial := CN[1].Initial
CN.Terminal := CN[CN.ListLength].Terminal
vary i over Nat
    satisfying i < CN.ListLength
        CN[i].Terminal.NT := CN[i + 1].Initial
endvary
```

These two definitions are a simple solution, which works fine for many examples. Other solutions can be defined with a generic Montage for lists.

In the graphs of Montages, a list node is represented by a box, which is marked in the right–top corner with the keyword LIST. A second box or circle within the LIST–box represents the single elements in the list. An arrow from a node within a LIST–box corresponds therefore to a family of arrows from all of the elements in that list, whereas an arrow to the list-box itself has the same semantics as normal arrows.

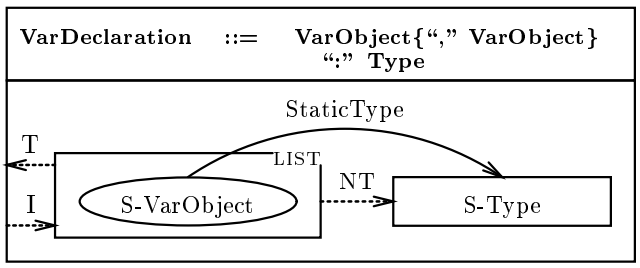


Figure 4: A variable declaration Montage

As example we take a Montage for a variable declaration, as pictured in fig. 4. The production rule of that Montage generates a list of VarObject-descendants and a node labeled with Type. The single *StaticType*-arrow in the Montage specifies a family of data flow arrows, one from each variable object to the type-node. Please note that the action of the list node links all variable objects sequentially with an NT arrow.

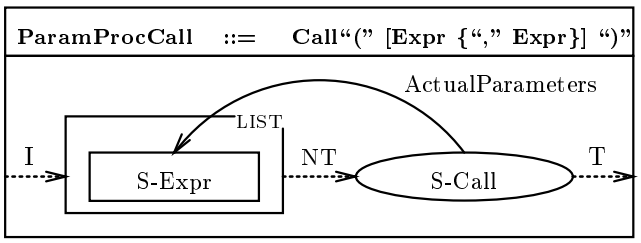


Figure 5: A Montage for a parameterized procedure call

Experience showed that often binary functions which link a leaf with all elements of a list are used. The second argument of such a function is the position in the list. We allow to define this type of binary function graphically by means of a data arrow going into a list box. An example where this is needed are the actual parameters of a procedure call (see fig. 5). The ActualParameters arrow in the Montage defines a binary function ActualParameters which maps a call task c and a position n to the n-th actual parameter of c. For shortness we do not give the third and forth part of a procedure call Montage and refer the reader to [KP97].

4.2 Parallel and Non-Deterministic Evaluation Order

In most programming languages, the order in which arguments of expressions are evaluated is not defined. In some, e.g. ANSI-C, it is even not defined that the arguments must be evaluated sequentially. A correct specification must thus be equally abstract and cannot just fix one of the allowed evaluation orders. In the Montages framework this situation can be modeled using the same technique as for the sequentialized graph traversal (3): all tasks of a program are executed in parallel, and a relation is used to sequentialize the execution partially.

In order to do so, the initial leaf must be generalized to a set of initial leaves. The semantics of a control flow arrow defines then a relation rather than a function. For each control arrow this relation is established from the single terminal leaf of the source to each initial leaf of the target. Instead of the abstract program counter CT, there is a set *ToDo* of tasks which have still to be executed. Non-determinism can be modeled by an external function which prevents certain tasks in the ToDo set from being executed immediately. The dynamic behavior of this function would determine whether all arguments are evaluated at once, in some some sequential order, or in a mixed form, e.g. not allowing more than a certain number of additions to be done in parallel.

Multiple initial leaves can be specified graphically by several I-arrows. The semantics of a Montage with several I-arrows is, that the initial leaves are the union of all initial leaves of all descendants pointed by an I-arrow. The terminal leaf remains the terminal leaf of the descendant pointed by the T-arrow. As example we show how the graphical part of the Sum Montage looks in the parallel case (fig. 6). Please note that the other parts of the Montage are almost identical with the Montage given in fig. 3, which shows good qualities in extending behavior. Detailed solutions to the problems related with parallelism, e.g. concurrent recursive procedure calls, and a discussion why it is important to abstract from a concrete evaluation order are given in [HK97]. Of course the techniques presented there can be used for the specification of parallel and concurrent programming languages as well. It might be worth noting, that the resulting control and data flow graphs correspond to those widely used in the emerging area of Hardware/Software Co-Design [DMS96]. We plan to investigate whether Montages are suited as front-end for the numerous code generation and hardware synthesis techniques [GDWL92] which are based on control and data flow graphs.

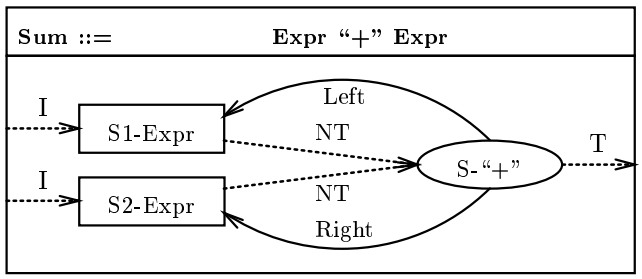


Figure 6: A Sum Montages where the arguments may be evaluated in parallel

4.3 Reuse of Montages

Although Montages are very compact, there is a large potential for reuse in the resulting specifications. We showed as example the Sum Montages. Most of that Montage can be reused for all binary expressions in a language. In C there are for instance ten different productions generating binary expressions. A mechanism for the reuse of the given Montage is therefore needed.

Originally we did not plan to design Montages in an object oriented way, but it turned out that they are object oriented by nature. Montages can be seen as class specifications which defines fields, initialization and methods for the tasks in a program. Among the fields are the control and data flow functions and their initialization is defined by the graph of montages. The

rest of the fields are dynamic functions, like the Value field of Expressions. Most Montages have only one method, which is implicitly called if control is passed to the task. This method corresponds to the dynamics semantics rule of the Montage. An example of tasks with more than one method are types, which provide a copy and a create method [KP97]. An inheritance relation between Montages is defined straightforward based on the view of them as classes.

It is interesting how in the case of Oberon there is a small number of classes, e.g. statements, expressions, designators, variable, and types, which form the inheritance hierarchy shown in figure 7. Each construct of Oberon can be modeled as a class inherited from one of the classes in figure 7. The specification of Oberon in [Kut97] is structured along these classes. We hope that it will be possible to derive from future experience a general theory of Montages classes for programming languages. In such a theory each existing language could be described as an instance of the general theory.

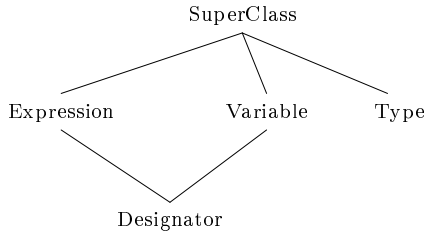


Figure 7: A class hierarchy of Montages of Oberon

The correspondence between Montages and classes in object oriented frameworks means we can adapt the well developed methodology of object oriented design to the Montages framework. On the other hand we can combine these informal methods with the formality of Montages and will get a new powerful methodology. Whether parts of this approach can be used for general object oriented design will be investigated in future research.

5 Related Works

Denotational semantics has been regarded as the most promising approach for the semantic description of programming languages. But its problems with the pragmatical side of language design have been discovered already in case studies of the scale of Pascal and C (see for instance [Sch86]). Information hiding, object orientedness and complex name analysis are not covered because of the global visibility of the definition of semantic domains throughout a denotation description. Moreover domain definitions often need to be changed when extending the language with unforeseen constructs, for instance a change from the direct style to the continuation style when adding *gotos* [Mos96]. To cite Abramsky " ... once languages with features beyond the purely functional are considered, the appropriateness of modeling programs by functions is increasingly open to question. Neither concurrency nor 'advanced' imperative features have been captured denotationally in a fully convincing fashion." [Abr96]

Other researches have been carried out because of the above considerations about pragmatics. In particular, it is worth mentioning Action semantics [Mos96], which is an *initial–algebra semantics* [GTWW77], based on Mosses' unified algebras. Action semantics retained some denotational semantics features, i.e. context–free grammars for defining abstract–syntax trees, and the use of Horn clauses to give inductive definition of compositional semantic functions. The main semantic entities are actions, which are specified by means of the action notation. To mention Mosses " ... the current structural operational semantics of action notation is not easy to modify; alternative forms of operational semantics, such as evolving–algebra semantics, might be preferable in that respect." [Mos96]

Another universal meta–language for all aspects of programming languages is ASF+SDF [vDHK96]. It is an initial–algebra approach and specifies the static and dynamic semantics by means of positive conditional equations. As all the initial–algebra based formalisms they are forced to remain under the expressiveness of the logic of Horn clauses, i.e. positive conditional equations, otherwise the existence of the initial model is not guaranteed and the syntax cannot be mapped in an unambiguous way to the semantics because of non existence of the universal homomorphism. In this respect, Montages are much more expressive since they make use of the full first–order logic for the static semantics predicates.

An approach with the same ambitious goal are Kahn's Natural Semantics [Kah87] which are directly based on Natural Deduction. For somebody knowing mathematical logic, Natural Semantics are pretty intuitive and we used it for the dynamic semantics of Oberon [Kut96]. Although we succeeded due to the excellent tool support by Centaur [BCD+87], the result was much longer and more complex then the GM counterpart given in [Kut97], since one has to carry around all the state information in the case of Natural Semantics.

Although attribute grammars [Knu68] are not designed to specify all aspects of languages, it's worth noting that the the solution for the static aspects of our approach has some similarities with attribute grammars. Although in certain cases they may be executed

very efficiently, we preferred not to use them for the following reasons: using GMs we have the same formalism for all parts of the specification; and as shown in [Ode89] attribute grammars tend to be very long if applied to real programming languages.

Using GMs for dynamic semantics, the work in [PH94] defines a framework comparable to ours. Although it has different aims, namely efficient execution. For the static part, it proposed occurrence algebras which integrate term algebras and context free grammars by providing terms for all nodes of all possible derivation trees. This allows such an approach to define all static aspects of the language in a functional algebraic system, which is supported by the MAX tool. In any case, the additional mathematical machinery must be hidden from the user and Montages might be well suited for that task.

6 Conclusions

In this paper we presented a novel approach to cope with specifications of all aspects of programming languages. Expressive yet intelligible descriptions of language constructs and of complex features together with ease of maintenance were sought. In this respect, the well known GMs have already attracted attention.

The classical use of GMs abstracts from the static semantics and assumes the result of a static analysis in order to define the dynamic semantics. The main criticisms against such an approach is that the static analysis is not formalized. An exception is the work on Occam [BD96]. Unfortunately the solution presented there allows not for the definition of the static semantics. Montages solves the problem using control and data flow graphs and at the same time allows one to give a very compact definition of static semantics as full first–order predicates. At the same time Montages retain the advantages of GMs.

Experience in scaling up both basic GMs and Montages for large case studies such as the specification of SQL [DiF97] and Oberon [Kut97, KP97] showed some important advantages of Montages with respect to basic GMs:

- the readability and comprehension of specifications improved drastically since the specification is arranged in capsules of behavior according to the rules of the context–free grammar,
- the maintenance of the whole specification is much easier since the behavior can be easily localized and eventually modified according to requirement changes,
- starting from the static semantics and analysis allows one to have a better comprehension of the dynamic semantics,
- the specification process requires less time since
  - the designer is driven by the structure of the Montages,
  - Montages represents a sort of factorization of behavior which can be reused in the definition of other languages in a off–the–shelf fashion or via refinement.

As pointed out in [Bör95] GMs are deliberately not imposing any particular calculus. The ease of abstraction makes GMs very suitable for different application domains and for each of them it makes sense to have a different verification system with its own assumptions. Moreover, one is able to do complex mathematical proofs directly in the GM framework, as shown in [Bör95, BD96, BR95]. The translation of a GM model in other formalisms is interesting if tools are available with powerful verification capabilities.

Montages are formal descriptions of programming languages with an higher intelligibility than usual semantics descriptions. This is mainly due to the use of visual elements and the elaborated structuring mechanism.

Future case studies will show whether Montages are not only suited for the specification of programming languages, but for other application domains as well, such as databases, hardware languages and communication protocols.

Acknowledgments We gratefully acknowledge Yuri Gurevich, David Espinosa, Egon Börger, Daniel Schweizer, and Jim Huggins for their constructive comments on early drafts of the paper. Thanks goes to Richard Waldinger and Chuck Wallace who helped us with the writing; Richard proposed the name Montages.

References

[Abr96] S. Abramsky. Semantics of Interaction. In *Trees in Algebra and Programming – CAAP'96, 21st Int. Coll.*, volume 1059 of *LNCS*, page 1. Springer Verlag, 1996.

[Anl97] M. Anlauff. The semantics of the object–oriented programming language sather. Technical report, International Computer Science Institute, Berkeley, 1997. In preparation.

[ASU85] A. Aho, R. Sethi, and J. Ullman. *Compilers, Principles, Techniques, and Tools.* Addison–Wesley, 1985.

[BCD+87] P. Borra, D. Clement, T. Despeyroux, J. Incerpi, G. Kahn, B. Lang, and V. Pascual. CENTAUR: The System. Technical Report 777, INRIA, Sophia Antipolis, 1987.

[BD96] E. Börger and I. Durdanović. Correctness of Compiling Occam to Transputer Code. *Computer Journal*, 39(1):52 – 92, 1996.

[BGM95] E. Börger, U. Gläser, and W. Mueller. Formal Definition of an Abstract VHDL'93 Simulator by EA–machines. In *Semantics of VHDL*, volume 307 of *The Kluwer International Series in Engineering and Computer Science*. Kluwer, 1995.

[Bör95] E. Börger. Why Use Evolving Algebras for Hardware and Software Engineering. In *SOFTSEM'95 22nd Seminar on Current Trends in Theory and Practice of Informatics*, volume 1012 of *LNCS*, pages 236 – 271. Springer Verlag, 1995.

[BR95] E. Börger and D. Rosenzweig. *The WAM - Definition and Compiler Correctness*, chapter 2, pages 20 – 90. Series in Computer Science and Artificial Intelligence. Elsevier Science B.V.North Holland, 1995.

[DCIG93] G. Del Castillo, Durdanović I., and U. Glässer. *An Evolving Algebra Abstract Machine*, volume 1092 of *LNCS*, pages 191 – 214. Springer Verlag, 1993.

[DiF97] B. DiFranco. Semantica Statica e Dinamica di SQL mediante i Montaggi. Master's thesis, Universitá di L'Aquila, 1997. (in italian), in preparation.

[DMS96] G. De Micheli and M. Sami, editors. *Hardware/Software Co-Design*, volume 310 of *NATO ASI*. Kluwer Academic, 1996.

[GDWL92] D.D. Gajski, N.D. Dutt, A.C-H Wu, and S. Y-L Lin. *High-Level Synthesis*. Kluwer Academic, 1992.

[GH93] Y. Gurevich and J.K. Huggins. *The Semantics of the C Programming Language*, volume 702 of *LNCS*, pages 274 – 308. Springer Verlag, 1993.

[GKS90] G. Gottlob, G. Kappel, and M. Schrefl. Semantics of Object–Oriented Data Models – The Evolving Algebra Approach. In *Next Generation Information System Technology, 1st International East/West Database Workshop*, volume 504 of *LNCS*, pages 144 – 160. Springer Verlag, 1990.

[GR94] P. Glavan and D. Rosenzweig. Evolving algebra model of programming language semantics. In *Technology and Foundations Information Processing '94*, volume A–51, pages 416 – 422. IFIP Transactions A (Computer Science and Technology), 1994.

[GTWW77] J.A. Goguen, J.W. Thatcher, E.G. Wagner, and J.B. Wright. Initial algebras semantics and continuous algebras. *J.ACM*, (24):68 – 95, 1977.

[Gur95] Y. Gurevich. Evolving Algebras 1993: Lipari Guide. In E. Börger, editor, *Specification and Validation Methods*. Oxford University Press, 1995.

[HK97] J.K. Huggins and P.W. Kutter. Abstraction from evaluation order and sequentiality. in preparation, 1997.

[Kah87] G. Kahn. Natural Semantics. In *Proceedings of the Symp. on Theoretical Aspects of Computer Science, Passau, Germany*, 1987.

[Knu68] D.E. Knuth. Semantics of Context–Free Languages. *Math. Systems Theory*, 2(2):127 – 146, 1968.

[KP96] P.W. Kutter and A. Pierantonio. Montages: Unified static and dynamic semantics of programming languages. Technical Report 118, Dip. Matematica Pura ed Applicata, Universitá di L'Aquila, July 1996.

[KP97] P.W. Kutter and A. Pierantonio. Montages: Towards the semantics of realistic programming languages. Technical Report 126, Dip. Matematica Pura ed Applicata, Universitá di L'Aquila, Jan 1997.

[Kut96] P.W. Kutter. Executable Specification of Oberon Using Natural Semantics. Term Work, ETH Zürich, implementation on the Centaur System [BCD+87], 1996.

[Kut97] P.W. Kutter. Dynamic Semantics of the Programming Language Oberon. Tik-report, ETH Zürich, 1997.

[Mos96] P. Mosses. Theory and Practice of Action Semantics. In *MFCS'96, 21st International Symposium*, volume 1113 of *LNCS*, pages 37 – 61. Springer Verlag, 1996.

[Ode89] M. Odersky. *A New Approach to Formal Language Definition and its Application to Oberon*. PhD thesis, ETH Zürich, 1989.

[PH94] A. Poetzsch-Heffter. Developing Efficient Interpreters Based on Formal Language Specifications. In *Compiler Construction*, volume 786 of *Lecture Notes in Computer Science*, pages 233 – 247. Springer–Verlag, 1994.

[Sch86] D.A. Schmidt. *Denotational Semantics: A Methodology for Language Development*. Allyn & Bacon, 1986.

[vDHK96] A. van Deursen, J. Heering, and P. Klint, editors. *Language Prototyping – An Algebraic Approach*, volume 5 of *AMAST Series in Computing*. World Scientific, 1996.

[Wal94] C. Wallace. The Semantics of the C++ Programming Language. In E. Börger, editor, *Specification and Validation Methods*, pages 131 – 164. Oxford University Press, 1994.