Compiler Generation Using HACS*

Kristoffer H. Rose
Two Sigma Investments/New York University

December 17, 2014

Abstract

Higher-order Attribute Contraction Schemes—or HACS—is a language for programming compilers. With HACS it is possible to create a fully functional compiler from a single source file. This document explains how to get HACS up and running, and walks through the code of a simple example with each of the main stages of a compiler in HACS: lexical analysis, syntax analysis, semantic analysis, and code generation.

1 Introduction

HACS abbreviates *Higher-order Attribute Contraction Schemes*, which is a formal system for symbolic rewriting extended with programming idioms commonly used when coding compilers. HACS is developed as a front-end to the CRSX higher-order rewriting engine [5].

A compiler written in HACS consists of a single *specification file* with a series of formal sections, each corresponding to a stage of the compiler. Each section is written in a formal style suitable for that stage of the compiler. Specifically, HACS supports the following notations:

Regular Expressions. Used to describe how an input text is partitioned into *tokens*. The regular expressions of HACS follows common conventions [1]. Section 3 gives details of this notation.

Context Free Grammars. HACS uses a form of BNF [4] *context free grammars*; the notation has been tweaked to look more like *templates* to allow for reuse of the notation in rewrite rules later. HACS includes simple mechanisms for automatic resolution of operator precedence and productions using immediate recursion such that the transformation from token stream to abstract syntax can be formalized. Details in Section 4.

Recursive Translation Schemes. Simple translations in general, and code generation in particular, are traditionally achieved by *recursive translation* from one abstract form to another. HACS includes special notations for defining such translations, as well as a mechanism for defining auxiliary so-called "semantic sorts," detailed in Section 5.

*This version describes HACS β version 0.9.16 released for use at NYU. This printing December 17, 2014.

Attribute Grammars. Analyses can be described with attribute grammars in the style of *Syntax-Directed Definitions* [1], originally introduced as *attribute grammars* [3], which describe how properties propagate through the abstract syntax tree. Section 6 details how the basic propagation rules work for synthesized attributes, Section 7 explains how inherited atributes are integrated such that full syntax-directed definitions can be encoded.

In the remainder of this document we introduce the most important features of the HACS language by explaining the relevant parts of the included *First.hx* example, inspired by [1, Figure 1.7], as well as several other minor examples. We first show in Section 2 how to install HACS and run the example, before we go on to how to write specifications. We explain lexical analysis in Section 3, syntax analysis in Section 4, basic semantic sorts and recursive translation schemes in Section 5, bottom-up semantic analysis in Section 6, and general syntax-directed definitions in Section 7. In Section 8 we explain how primitive values can be manipulated, and in Section 9 we give several examples of how everything can be combined.

Appendix A has a reference manual, Appendix B explains some of the (still) cryptic error messages, and Appendix C lists some current limitations.

2 Getting Started

In this section we walk through the steps for getting a functional HACS installation on your computer.[1]

2.1 Requirements. To run the HACS examples here you need a *nix system (including a shell and the usual utilities) with these common programs: a Java development environment (at least Java 1.6 SE SDK, with java and javac commands); and a standard *nix development setup including GNU Make and a C99 compiler. In addition, the setup process needs internet access to retrieve the CRSX base system [5], JavaCC parser generator [6], and *icu* Unicode C libraries [7].

2.2 Commands (install HACS)**.** Retrieve the *hacs-0.9.16.zip* archive, extract it to a new directory, and install it, for example with the following commands:[2]

```
energon1[~]$ wget http://crsx.org/hacs-0.9.16.zip
energon1[~]$ unzip hacs-0.9.16.zip
energon1[~]$ cd hacs
energon1[hacs]$ make FROMZIP=1 install install-support
```

These commands will need Internet access, using the wget command to retrieve support libraries.[3] The above command will install HACS in a *.hacs* subdirectory of your home directory. You can change this with the option `prefix=...` to (all uses of) *make*; if so then make sure to replace occurrences of `$HOME/.hacs` everywhere below with your chosen directory.

The main *make* command will take some time (of the order of 3 minutes) but should end without error.

The following command makes the main *hacs* command available for use:

```
energon1[hacs]$ alias hacs=$HOME/.hacs/bin/hacs
```

[1] HACS is still a β release so please report any problems with this procedure to *hacs-bugs@crsx.org*.

[2] User input is blue.

[3] Specifically, HACS needs the CRSX system, JavaCC parser generator, and ICU4C Unicode library. The setup is documented in *src/Env.mk*.

(It may be worth including this command in your setup, or including the `$HOME/.hacs/bin` directory in your `$PATH`.)

Please check that your new installation works with these commands:

```
energon1[hacs]$ cd
energon1[~]$ mkdir myfirst
energon1[~]$ cd myfirst
energon1[~]$ cp $HOME/.hacs/share/doc/hacs/examples/First.hx .
energon1[~]$ $HOME/.hacs/bin/hacs First.hx
energon1[~]$ ./First.run --scheme=Compile \
           --term="{initial := 1; rate := 1.0; position := initial + rate * 60;}"
  LDF T_2,  #1
    STF initial, T_2
    LDF T_2_37,  #1.0
    STF rate, T_2_37
    LDF T_3,  initial
    LDF T_3_42,  rate
    LDF T_4,  #60
    MULF  T_4_71 ,  T_3_42 ,  T_4
    ADDF  T_2_49 ,  T_3 ,  T_4_71
    STF position, T_2_49
```

Congratulations—you just built your first compiler![4]

2.3 Example (module wrapper)**.** The source file for the *First.hx* file used in the example above has the structure

```
/* Top comment. */
module org.crsx.hacs.samples.First
{
  // Remark.
  Lexical Analysis (Section 3)
  Syntax Analysis (Section 4)
  Semantic Analysis (Sections 5, 6 and 7)
  Code Generator (Section 9)
  Main (Section 5)
}
```

Notice that HACS permits C/Java style comments.

2.4 Notation (special Unicode characters)**.** HACS uses a number of special symbols from the standard Unicode repertoire of characters, shown in Table 1

3 Lexical Analysis

Lexical analysis is the process of splitting the input text into tokens. HACS uses a rather standard variation of *regular expressions* for this.

3.1 Example (tokens and white space)**.** Here is a HACS fragment for setting up the concrete syntax of integers, basic floating point numbers, identifiers, and white space, for use by a simple language:

[4]Please do not mind the spacing – that is how HACS prints in its present state.

Glyph	*Code Point*	*Character*
¬	U+00AC	logical negation sign
×	U+00D7	multiplication sign
÷	U+00F7	multiplication sign
¶	U+00B6	paragraph sign
↑	U+2191	upwards arrow
→	U+2192	rightwards arrow
↓	U+2193	downwards arrow
⟦	U+27E6	mathematical left white square bracket
⟧	U+27E7	mathematical right white square bracket
⟨	U+27E8	mathematical left angle bracket
⟩	U+27E9	mathematical right angle bracket

Table 1: Unicode special characters used by HACS.

```
// White space convention.                                    1
space [ \t\n] ;                                               2

// Basic nonterminals.                                        4
token INT      | ⟨DIGIT⟩+ ;                                   5
token FLOAT    | ⟨DIGIT⟩* "." ⟨DIGIT⟩+ ;                      6
token ID       | ⟨LOWER⟩+ ('_'? ⟨INT⟩)* ;                     7

// Special categories of letters.                             9
token fragment DIGIT | [0−9] ;                                10
token fragment LOWER | [a−z] ;                                11
```

The example illustrates the following particulars of HACS lexical expressions:

- Declarations generally start with a keyword or two and are terminated by a ; (semicolon).
- token declarations in particular have the token keyword followed by a regular expression between a | (vertical bar) and a ; (semicolon). It defines the token as a *non-terminal* that can be used in syntax productions described in the next section.
- A regular expressions is a sequence of units, corresponding to the concatenation of sequences of characters that match each one. Each unit can be a *character class* such as [a−z], which matches a single character in the indicated range (or, more generally, a sequence of individual characters and ranges), a *string* such as ".", or a *reference* to a token or fragment such as ⟨Lower⟩, enclosed in the special Unicode mathematical angle brackets (see Table 1).
- A token fragment declaration means that the defined token can only be used in other token declarations, and not as grammar non-terminal in syntax productions.
- Every regular expression component can be followed by a repetition marker ?, +, or ∗, and regular expressions can be *grouped* with parentheses.
- The regular expression for white space is setup by space followed by the regular expression of what to skip – here spaces, tabs, and newlines, where HACS uses backslash for escaping in character classes with usual C-style language escapes.

In addition, we have followed the convention of naming proper grammar terminals with ALL-CAPS names, like INT, so they are easy to distinguish from non-terminals below.

Notice that while it is possible to make every keyword of your language into a named token in this way, this is not necessary, as keywords can be given as literals in syntax productions, covered in the next section.

3.2 Commands (lexical analysis)**.** The fragment above is part of *First.run* from Section 1, which can thus be used as a lexical analyzer. This is achieved by passing the *First.run* command two arguments: a *token sort* and a *token term*.[5] Execution proceeds by parsing the string following the syntax of the token. We can, for example, check the lexical analysis of a number:

```
$ ./First.run --sort=FLOAT --term=34.56
34.56
```

If there is an error, the lexical analyzer will inform us of this:

```
$ ./First.run --sort=INT --term=34.56
Exception in thread "main" java.lang.RuntimeException: net.sf.crsx.CRSException:
  Encountered " <T_FLOAT> "34.56 "" at line 1, column 1.
Was expecting:
    <T_INT> ...
```

(where the trail of Java exceptions has been truncated: the important information is in the first few lines).

4 Syntax Analysis

Once we have tokens, we can use HACS to program a syntax analysis with a grammar that specifies how the input text is decomposed according to a *concrete syntax* and how the desired *abstract syntax tree* (AST) is constructed from that. Notice that HACS does not provide a "parse tree" in the traditional sense, *i.e.*, a tree that represents the full concrete syntax parse: only the AST is built. Grammars are structured following the *sorts* of AST nodes, with concrete syntax details managed through annotations and "syntactic sugar" declarations.

4.1 Example. Here is another extract from our *First.hx* example, with a small syntax analysis, or grammar. Our small example source language merely has blocks, assignment statements, and a few forms of expression, like so:

```
// Main program construct.                                              1
main sort Stat  | ⟦ ⟨Name⟩ := ⟨Exp⟩ ; ⟨Stat⟩ ⟧                          2
                | ⟦ { ⟨Stat⟩ } ⟨Stat⟩ ⟧                                  3
                | ⟦⟧ ;                                                   4

sort Exp | ⟦ ⟨Exp⟩ + ⟨Exp@1⟩ ⟧                                           6
         | ⟦ ⟨Exp@1⟩ * ⟨Exp@2⟩ ⟧@1                                      7
         | ⟦ ⟨INT⟩ ⟧@2                                                  8
         | ⟦ ⟨FLOAT⟩ ⟧@2                                                9
         | ⟦ ⟨Name⟩ ⟧@2                                                 10
         | sugar ⟦ ( ⟨Exp#⟩ ) ⟧@2 →Exp# ;                               11
```

[5] The command has more options that we shall introduce as we need them.

```
sort Name | symbol ⟦⟨ID⟩⟧ ;                                                    13
```

The grammar structures the input as three sorts: Stat for statements, Exp for expressions, and Name for names (which we shall need later for symbol tables).

The example grammar above captures the HACS version of several standard parsing notions:

Literal syntax is indicated by the double "syntax brackets," ⟦...⟧. Text inside ⟦...⟧ consists of three things only: space, literal character "words," and references to non-terminals and predefined tokens inside (nested) ⟨...⟩. In this way, literal syntax is similar to macro notation or "quasi-quotation" of other programming languages.

Syntactic sugar is represented by the sugar part of the Exp sort declaration, which states that the parser should accept an Exp in parenthesis, identified as #, and replace it with just that same Exp, indicated by the →Exp# part. This avoids any need to think of parentheses in the generated AST.

Precedence rules are represented by the @-annotations, which assign precedence and associativity to each operator. Thus the same Exp language would be recognized by something like the following concrete HACS specification (where we have also made the sugar concrete):

```
sort Exp  | ⟦ ⟨Exp0⟩ ⟧;
sort Exp0 | ⟦ ⟨Exp0⟩ + ⟨Exp1⟩ ⟧| ⟦⟨Exp1⟩ ⟧;
sort Exp1 | ⟦ ⟨Exp1⟩ * ⟨Exp2⟩ ⟧| ⟦ ⟨Exp2⟩ ⟧;
sort Exp2 | ⟦ ⟨Int⟩ ⟧ | ⟦ ⟨Float⟩ ⟧ | ⟦ ⟨Name⟩ ⟧| ⟦ ( ⟨Exp⟩ ) ⟧ ;
```

However, this grammar generates a different result tree, where the nodes have the four different sorts used instead of all being of the single Exp sort that the precedence annotations make possible. The transformed system also illustrates how HACS deals with left recursion with @-annotations: each becomes an instance of *immediate left recursion*, which is eliminated automatically.

The precedence notation allows us to define one sort per "sort of abstract syntax tree node." This allows us to use a single kind of AST node to represent all the "levels" of expression, which helps the subsequent steps.

The notation is admittedly dense: this is intentional, as we generalize this very same notation to serve all the formalisms of the following sections. Here are the formal rules:

- Each sort is defined by a sort declaration followed by a number of *productions*, each introduced by a | (bar). (The first | corresponds to what is usually written "::=" or "→" in grammars.)
- Concrete syntax is enclosed in ⟦...⟧ ("double" or "white" brackets). Everything inside double brackets should be seen as *literal syntax*, even \ (backslash), *except* for HACS white space (corresponding to [\t\n\r]), which is ignored, and references in ⟨...⟩ (angle brackets), which are special.
- References to *nonterminals* (other productions) are wrapped in ⟨...⟩ (angle brackets).
- *Precedence* is indicated with @n, where higher numbers n designate higher (tighter) precedence. After every top-level ⟦⟧ and inside every ⟨⟩ there is a precedence, which defaults to @0.

- The special sugar declaration expresses that the concrete syntax can use parentheses to raise the precedence of the enclosed expression to 2: it is the first example of a *rewrite rule* with a $\rightarrow$ that we see, where we remark that the expression is marked # so we can use Exp# to the right of the $\rightarrow$ to indicate that the result of simplifying concrete syntax with parenthesis. (In fact the general rule is that when an $\rightarrow$ is used then all sort specifiers must be "disambiguated" with distinct markers like # or #5 in this way.)
- As a special case, A sort can be defined with a single symbol declaration. If so, then the actual syntax must refer to a single token, and that token must allow multiple trailing $\_n$ (underscore and count), which is added to any instance (this permits automatic symbol generation).

Of all these rules, the one thing that is unique to parsing is the precedence notation with @. When specifying a grammar then *every* subterm has a precedence, which determines how ambiguous terms should be parsed. So imagine that every $\langle\rangle$ contains a @ marker, defaulting to @0, and that every $[\![]\!]$ is terminated with a @ marker, again defaulting to @0.

Notice that HACS will do two things automatically:

1. Eliminate immediate left recursion, such as found in the example.
2. Split the productions into subproductions according to the precedence assignments.
3. Left factor the grammar, which means that productions within a sort may start with a common prefix.

However, this is *not* reflected in the generated trees: they will follow the grammar as specified, so you do not need to be aware that this conversion happens.

4.2 Commands. We can parse an expression from the command line:

```
$ ./First.run --sort=Exp --term="(2+(3*(4+5)))"
2 + 3 * ( 4 + 5 )
```

Notice that the printout differs slightly from the input term as it has been "resugared" from the AST with minimal insertion of parentheses.

5 Sorts and Recursive Translation Schemes

In this section we explain how basic algebraic structures and transformations are expressed in HACS.

5.1 Example. Another fragment of the *First.hx* example has the semantic sorts and operations that are used. For our toy language that just means the notion of a *type* with the way that types are "unified" to construct new types.

```
// Types to associate to AST nodes.                                    1
sort Type | Int | Float ;                                              2

// The Type sort includes a scheme for unifying two types.             4
| scheme Unif(Type,Type) ;                                             5
Unif(Int, Int) → Int;                                                  6
Unif(#1, Float) →Float;                                                7
Unif(Float, #2) →Float;                                                8
```

The code declares a new sort, Type, which is a *semantic* sort because it does not include any syntactic cases: all the possible values (as usual listed after leading |s) are simple *term structures* written without any ⟦⟧s. Structures are written with a leading "constructor," which should be a capitalized word (the same as sort names), optionally followed by some "arguments" in ()s, where the declaration gives the sort for each argument (here there are none).

The semantic sort also includes a scheme declaration for the Unif constructor, which must be followed by an argument list with two Type arguments. The scheme declaration is "instantiated" by *rules* of the form

pattern→replacement

which must specify for each possible shape of Unif-construction how it should be simplified by the scheme. Rules may include *parameters* in the form of "meta-variables" starting with # (hash), like #1, to designate "function arguments" that should be copied from the pattern to the replacement.

The rules for the Unif scheme above can, for example, be used to simplify a composite term as follows:

Unif(Unif(Int, Float), Int) → Unif(Float, Int) → Float

Note how overlaps are allowed but please do verify determinacy, *i.e.*, if a particular combination of arguments can be subjected to two rules then they should give the same result! In this example i happens because the term Unif(Float,Float) can be rewritten by both the rule in line 7 and 8, but it does not matter, because the result is the same.

5.2 Example (scheme over syntax)**.** The Unif scheme defined in the example is a simple example of a semantic recursive translation scheme, defined by rewrite rules. We are permitted to define such schemes over the syntactic sorts, as well. Here is, for example, code to extract the leftmost leaf expression from an Exp tree from Example 4.1:

```
sort Exp | scheme Leftmost(Exp) ;
Leftmost(⟦⟨Exp#1⟩ + ⟨Exp#2⟩⟧) →Leftmost(Exp#1) ;
Leftmost(⟦⟨Exp#2⟩ * ⟨Exp#3⟩⟧) →Leftmost(Exp#1) ;
Leftmost(⟦⟨INT#⟩⟧) →⟦⟨INT#⟩⟧ ;
Leftmost(⟦⟨FLOAT#⟩⟧) →⟦⟨FLOAT#⟩⟧ ;
Leftmost(⟦⟨Name#⟩⟧) →⟦⟨Name#⟩⟧ ;
```

Notice how:

- We first set the current sort to Exp and then add a scheme called Leftmost that takes one argument of the syntactic Exp sort.
- There is precisely one rule with a pattern applying Leftmost to each non-sugar production for Exp from Example 4.1.
- Each non-terminal reference has the non-terminal name followed by a $\#n$ marker to identify the subexpression of that non-terminal sort for use on the right side of the →.
- Each rule rewrites an Exp expression to another Exp expression, either by recursively invoking the defined Leftmost scheme on a smaller part of the term or by returning the term itself.
- We *have* to write ⟦⟨Name#⟩⟧ rather than just Name# or # in the last rules because the form Name# describes something of Name sort, not Exp sort. Indeed in the last rule, Name# stands for the subterm of the Exp expression that is a Name.

5.3 Commands (invoke scheme)**.** The Leftmost scheme above is also included in *First.hx*. Since it operates on a syntactic expression, we can invoke the Leftmost scheme from the command line as follows:

```
$ ./First.run --scheme=Leftmost --sort=Exp --term="(2*3)+4"
 2
```

We specify the sort of our input expression here because Leftmost takes an Exp argument, which is different from the main Stat sort.

Note that we cannot meaningfully invoke the Unif scheme from the command line because there is no user syntax for types in our example!

5.4 Example (syntactic scheme)**.** We can even go all the way and define a purely "syntactic scheme." This is best explained with an example. Consider the list sort

```
sort List | ⟦ ⟨Elem⟩ ⟨List⟩ ⟧ | ⟦⟧ ;
```

If we want to allow appending and flattening of lists, then we can define

```
sort List | scheme ⟦{ ⟨List⟩ } ⟨List⟩ ⟧ ;
⟦ { ⟨Elem#1⟩ ⟨List#2⟩ } ⟨List#3⟩ ⟧→⟦⟨Elem#1⟩ { ⟨List#2⟩ } ⟨List#3⟩ ⟧;
⟦ { } ⟨List#⟩ ⟧ → # ;
```

Notice that:

- The two rules differ on the content of the braces, and are clearly designed to fully eliminate all braces used: this is essential, and we say that the scheme should be *complete*.
- This is very like sugar productions except we can have more than one rule for a given construct – sugar is limited to a single rule that cannot depend on the inner structure of the construct.
- If you work with synthesized attributes (explained in the next section) then be aware that syntactic rules such as the ones presented this one will mess with any synthetic attributes.

5.5 Example. The *First.hx* example defines the Compile scheme as a top level function to be used from the command line. This looks as follows:

```
sort AProgr | scheme Compile(Stat);                          1
Compile(#) →⟦CG ICG TA ⟨Stat#⟩ ⟧;                            2
```

The Compile scheme reflects how our compiler is structured, and in particular that the input is a Stat and the output an AProgr, which stands for "assembly program." (You will see later what the right side of the → here means.) This is the reason for the `--scheme=Compile` option we invoked back in the getting started section. Such wrapper raw schemes must have a single argument.

6 Collecting Information

HACS has special support for assembling information in a "bottom-up" manner, corresponding to how *synthetic attributes* are used in syntax-directed definitions (or attribute grammars). In this section we explain *how* you convert any SDD synthetic attribute definition into a HACS one.

Consider the following single definition of the synthesized attribute t for expressions E:

PRODUCTION	SEMANTIC RULES
$E \to E_1 + E_2$	$E.t = \text{Unif}(E_1.t, E_2.t)$

(E1)

The rule is "S-attributed" because it exclusively relies on synthesized attributes. This allows us to express it directly in HACS as follows.

1. The first thing to do is declare the attribute and associate it with the E sort.

   ```
   attribute ↑t(Type);
   sort E | ↑t ;
   ```

   the ↑ indicates "synthesized" because the attribute moves "up" in the tree. The declaration of the attribute indicates with (Type) that the *value* of the synthesized attribute is a Type. Attributes are always named with lower case names.

2. The second thing to do is copy the corresponding syntax production but omit any @-markings and add a unique $\#n$ disambiguation mark to each production, essentially "upgrading" each nonterminal reference to a meta-variable. Since (E1) is based on the alternative

   ```
   sort E   | ⟦ ⟨E@1⟩ + ⟨E@2⟩ ⟧@1
   ```

   similarly to Example 4.1, we start with

   ```
   ⟦ ⟨E#1⟩ + ⟨E#2⟩ ⟧
   ```

   where we have used the subscripts from (E1) as #-disambiguation marks.

3. Next add in *synthesis patterns* for the attributes we are reading. Each attribute reference like $E_1.t$ becomes a pattern like ⟨E#1 ↑t(#t1)⟩, where the meta-variables like #t1 should each be unique. For our example, this gives us

   ```
   ⟦ ⟨E#1 ↑t(#t1)⟩ + ⟨E#2 ↑t(#t2)⟩ ⟧
   ```

   which sets up #t1 and #t2 as synonyms for $E_1.t$ and $E_2.t$, respectively.

4. Finally, add in the actual synthesized attribute, using the same kind of pattern at the *end* of the rule (and add a ;), and we get

   ```
   ⟦ ⟨E#1 ↑t(#t1)⟩ + ⟨E#2 ↑t(#t2)⟩ ⟧↑t(Unif(#t1,#t2)) ;
   ```

   This is read "When considering an E (the current sort) which has the shape $[\![\langle E\rangle + \langle E\rangle]\!]$ where furthermore the first expression has a value matching #t1 for the synthesized attribute t, and the second expression has a value matching #t2 for the synthesized attribute t, then the entire expression should be assigned the value Unif(#t1,#t2) for the synthesized attribute t."

6.1 Example. In Example 4.1, we presented the abstract syntax of the small language processed by *First.hx*. A type analysis of the expressions of the language excluding variables might look as follows as a standard SDD (syntax directed definition), where we use E for the Exp non-terminal, and one attribute: $E.t$ is the synthesized Type of the expression E. In the notations of [1], the SDD can be specified something like this:

PRODUCTION	SEMANTIC RULES
$E \to E_1 + E_2$	$E.t = \mathrm{Unif}(E_1.t, E_2.t)$
$\mid E_1 * E_2$	$E.t = \mathrm{Unif}(E_1.t, E_2.t)$
$\mid$ **int**	$E.t = \mathrm{Int}$
$\mid$ **float**	$E.t = \mathrm{Float}$

where we assume that Unif is defined as discussed in Example 5.1. We can convert this SDD to the following HACS (using the proper names for the sorts as actually found in *First.hx*):

```
attribute ↑t(Type);          // synthesized type                                    1

sort Exp | ↑t ;              // expressions have an associated synthesized type, E.t  3

// Synthesis rules for E.t.                                                         5
⟦ ⟨Exp#1 ↑t(#t1)⟩ + ⟨Exp#2 ↑t(#t2)⟩ ⟧↑t(Unif(#t1,#t2));                           6
⟦ ⟨Exp#1 ↑t(#t1)⟩ * ⟨Exp#2 ↑t(#t2)⟩ ⟧↑t(Unif(#t1,#t2));                           7
⟦ ⟨INT#⟩ ⟧↑t(Int);                                                                8
⟦ ⟨FLOAT#⟩ ⟧↑t(Float);                                                            9
```

Line 1 declares the value of the synthesized `t` attribute to be a Type. Line 3 associates the synthetic attribute `t` to the Exp sort: all synthetic attributes are associated with one or more abstract syntax sorts. The remaining lines 5–9 are *synthesis rules* that show for each form of Exp what the value should be, based on the values passed "up" from the subexpressions; these are generated mechanically as discussed above from the synthesis semantic rules.

Finally, note that if you have multiple synthetic attributes then a synthesis rule *only* adds one new attribute to the program construct in question, it does not remove any other attributes already set for it.

7 Full Syntax-Directed Definitions

In general, however, we wish to implement analyses that are more general than what can be achieved with S-attributed syntax-directed definitions: we also want to use *inherited* attributes. This sectionexplains how inherited attributes are implemented in HACS.

Consider the following two simple semantic rules:

PRODUCTION	SEMANTIC RULES	
$S \to \mathbf{name}_1 := E_2; S_3$	$E_2.e = S.e; S_3.e = \text{Extend}(S.e, \mathbf{name}_1.sym, E_2.t)$	(1)
$E \to E_1 + E_2$	$E_1.e = E.e; E_2.e = E.e$	(2)

where we furthermore have the property that th4 $E.t$ property cannot be synthesized until *after* the inherited attribute $E.e$ has been "spread" into the term (we shall see later why this is typical).

Here are the steps to follow to translate the inheritance to HACS.

1. The first thing to do is, again, to declare the attribute. Since $S.e$ and $E.e$ are a *map* from names to types, this is written as follows:

   ```
   attribute ↓e{Name:Type} ;
   ```

   The ↓ indicates an inherited attribute, and the {Name:Type} part declares the value of the attribute to be a mapping from values of Name sort to values of Type sort. (We'll assume that we still have the $E.t$ synthesized attribute defined as previously.) As above, attributes are always given lower case names.

   Note that only sorts with a symbol declaration can be used for keys of mappings: the mappings take the rôle of *symbol tables* from traditional compilers.

2. The second thing to do is to associate the inherited attribute to a *recursive scheme*, which will be responsible for propagating that inherited attribute over a values of a certain sort. This has to be done separately for each sort that e propagates over. Rule (2) is the simplest, so we take that first. The rule propagates the e attribute over E subexpressions, so we invent a scheme, Ee, which does that.

   sort E | scheme Ee(E) ↓e ;

   As can be seen, the scheme generates results of the E sort and also takes parameters of the E sort; in addition it *carries* the inherited attribute e.

3. We first observe that (2) operates on sums, which in our case have a syntax like this:

   sort E | ⟦ ⟨E@1⟩ + ⟨E@2⟩ ⟧@1 ;

   As before, we create a *pattern* that is equivalent to this, using the subscripts from (2):

   ⟦ ⟨E#1⟩ + ⟨E#2⟩ ⟧

4. Now *insert* the pattern into the scheme:

   Ee(⟦⟨E#1⟩ + ⟨E#2⟩⟧)

   This clearly respects the sort constraints we defined above, with Ee being applied to an E expression.

5. Since there are no complicated dependencies in (2), we are almost done: we just have to create a rule where we on the right side of the → *apply* the Ee scheme recursively to the subexpressions that should inherit the e attribute:

   Ee(⟦⟨E#1⟩ + ⟨E#2⟩⟧) →⟦⟨E Ee(#1)⟩ + ⟨E Ee(#2)⟩⟧ ;

   Notice that there is no explicit mention of the e attribute, only the *implicit* copying that follows from the use of the Ee scheme. The recursive arrangement of the Ee wrappers implies the two attribute equations $E_1.e = E.e$ and $E_2.e = E.e$ from (2).

6. The above rule implements (2), with one caveat: every time the e attribute is propagated through, a new expression is created, which thus loses all the synthesized properties it may have had. If we know that the transformation does not invalidate any of the already synthesized attributes, then we express that by augmenting the rule to

   Ee(⟦⟨E#1⟩ + ⟨E#2⟩⟧ ↑#syn) →⟦⟨E Ee(#1)⟩ + ⟨E Ee(#2)⟩⟧ ↑#syn ;

   which now explicitly declares that the new expression should keep all synthetic attributes (again named with a meta-variable so several sets can be preserved by a single rule).

7. Rule (1) is slightly more complicated, because the inherited attribute has non-trivial dependencies. We must know the dependency relationship of the attributes to devise a *recursive strategy* for the attribute evaluation. Recall that we have the following (realistic) dependency for (1): "The $E_2.t$ attribute cannot be computed until *after* $E_2.e$ has been instantiated (and recursively propagated)." In that case we have to evaluate (1) in two steps:

   (a) Do $E_2.e = S.e$, establishing the precondition for allowing the system to compute $E_2.t$.

   (b) When the system has computed $E_2.t$ then do $S_3.e = \mathrm{Extend}(S.e, \mathbf{name}_1.sym, E_2.t)$.

These two steps are achieved by having an extra carrier schemes:

```
sort S | scheme Se(S) ↓e | scheme SeB(S) ↓e ;
```

The first propagates into statements in the same way as for (2), except only into the first subterm, and *chains to the next stage*:

```
Se(⟦x := ⟨E#2⟩; ⟨S#3⟩⟧ ↑#syn) → SeB(⟦x := ⟨E Ee(#2)⟩; ⟨S#3⟩⟧ ↑#syn) ;
```

Notice how we invoke the Ee scheme to pass $E_2.e$ (the system understands that $S.e$ and $E_2.e$ refer to the same attribute), and how we do *not* wrap #3 in anything as nothing should be passed there; instead we just leave the extra top wrapper SeB to wait for when it can process.

Once the synthetic is satisfied, *i.e.*, once $E_2.t$ is computed, then the second stage can finishes the job, using the notation of the previous section, and proceed recursively on the appropriate subterm. However, it also needs to compute the Extend result, which is achieved with the following rule:

```
SeB(⟦x := ⟨E#2 ↑t(#t2)⟩; ⟨S#3⟩⟧ ↑#syn) →⟦x := ⟨E#2⟩; ⟨S Se(#3)↓e{⟦x⟧:#t2} ⟩⟧↑#syn ;
```

The last bit of notation is the way we call Se with an *extended* $S_3.e$ attribute: the notation Se(#3)↓e{⟦x⟧:#t2} means "call Se (passing e) on the statement #3 but use an e which has been extended with the mapping from x to #t2," which corresponds to the notation $S_3.e = \mathrm{Extend}(S.e, \mathbf{name}_1.sym, E_2.t)$ of the SDD.

Also note that because the Name sort is a symbol sort, we should *directly* use the variable x (which is a legal ID token) in the rules instead of ⟨Name#x⟩ or such. However, we still take care to distinguish a symbol that is part of the syntax, like x, from the other symbols that are part of the formalism, like S and Se: x is always enclosed in ⟦⟧s

PRODUCTION	SEMANTIC RULES	
$S \to \mathbf{name} := E_1; S_2$	$E_1.e = S.e; S_2.e = \mathrm{Extend}(S.e, \mathbf{name}.sym, E_1.t)$	(S1)
$\mid \{\ S_1\ \}\ S_2$	$S_1.e = S.e; S_2.e = S.e$	(S2)
$\mid \epsilon$		(S3)
$E \to E_1 + E_2$	$E_1.e = E.e; E_2.e = E.e; E.t = \mathrm{Unif}(E_1.t, E_2.t)$	(E1)
$\mid E_1 * E_2$	$E_1.e = E.e; E_2.e = E.e; E.t = \mathrm{Unif}(E_1.t, E_2.t)$	(E2)
$\mid \mathbf{int}$	$E.t = \mathrm{Int}$	(E3)
$\mid \mathbf{float}$	$E.t = \mathrm{Float}$	(E4)
$\mid \mathbf{name}$	$E.t = \mathrm{if}\ \mathrm{Defined}(E.e, \mathbf{name}.sym)$ $\quad \mathrm{then}\ \mathrm{Lookup}(E.e, \mathbf{name}.sym)$ $\quad \mathrm{else}\ \mathrm{TypeError}$	(E5)

Figure 1: SDD for type checking.

7.1 Example. An SDD for a simple type analysis can be implemented with two attributes (and using the usual convention that the SDD uses E and S where the HACS grammar has Exp and Stat):

```
sort Type | Int | Float | TypeError                                                          1
         | scheme Unif(Type,Type) ;                                                        2

Unif(Int, Int) → Int;                                                                     4
Unif(#t1, Float) →Float;                                                                  5
Unif(Float, #t2) →Float;                                                                  6
default Unif(#1,#2) →TypeError; // fall-back                                              7

attribute↑t(Type); // synthesized expression type                                         9
sort Exp | ↑t;                                                                            10

⟦ (⟨Exp#1 ↑t(#t1)⟩ + ⟨Exp#2 ↑t(#t2)⟩) ⟧↑t(Unif(#t1,#t2));                                 12
⟦ (⟨Exp#1 ↑t(#t1)⟩ * ⟨Exp#2 ↑t(#t2)⟩) ⟧↑t(Unif(#t1,#t2));                                 13
⟦ ⟨INT#⟩ ⟧↑t(Int);                                                                        14
⟦ ⟨FLOAT#⟩ ⟧↑t(Float);                                                                    15
// Missing case: variables – handled by Ee below.                                         16

attribute↓e{Name:Type}; // inherited type environment                                     18

sort Exp | scheme Ee(Exp) ↓e ; // propagates e over Exp                                   20

// These rules associate t attribute with variables (missing case above).                 22
Ee(⟦id⟧) ↓e{⟦id⟧ : #t} →⟦ id ⟧ ↑t(#t);                                                    23
Ee(⟦id⟧) ↓e{¬⟦id⟧} → error⟦Undefined identifier ⟨id⟩⟧;                                    24

Ee(⟦⟨Exp#1⟩ + ⟨Exp#2⟩⟧ ↑#syn) →⟦⟨Exp Ee(#1)⟩ + ⟨Exp Ee(#2)⟩ ⟧↑#syn ;                    26
Ee(⟦⟨Exp#1⟩ * ⟨Exp#2⟩⟧ ↑#syn) →⟦⟨Exp Ee(#1)⟩ * ⟨Exp Ee(#2)⟩ ⟧↑#syn ;                    27
Ee(⟦⟨INT#⟩⟧ ↑#syn) →⟦⟨INT#⟩⟧ ↑#syn ;                                                      28
Ee(⟦⟨FLOAT#⟩⟧ ↑#syn) →⟦⟨FLOAT#⟩⟧ ↑#syn ;                                                  29

sort Stat | scheme Se(Stat) ↓e ; // propagates e over Stat                                31

Se(⟦id := ⟨Exp#1⟩; ⟨Stat#2⟩⟧ ↑#syn) →SeB(⟦id := ⟨Exp Ee(#1)⟩; ⟨Stat#2⟩⟧ ↑#syn);            33
{                                                                                         34
  | scheme SeB(Stat) ↓e; // helper scheme for assignment after expression typeanalysis    35
  SeB(⟦id := ⟨Exp#1 ↑t(#t1)⟩; ⟨Stat#2⟩ ⟧↑#syn)                                           36
   → ⟦id := ⟨Exp#1⟩; ⟨Stat Se(#2) ↓e{⟦id⟧:#t1}⟩⟧ ↑#syn ;                                37
}                                                                                         38

Se (⟦ { ⟨Stat#1⟩ } ⟨Stat#2⟩ ⟧↑#syn) →⟦{ ⟨Stat Se(#1)⟩ } ⟨Stat Se(#2)⟩ ⟧↑#syn ;             40

Se (⟦ ⟧ ↑#syn) →⟦ ⟧ ↑#syn ;                                                               42
```

Figure 2: HACS code for type analysis.

Operators	*Explanation*
$+, -$	addition, subtraction
$\|, \hat{}, \backslash$	bitwise or, exclusive or, clear
$\times, \div, \%$	multiplication, divison, modulo
$\&$	bitwise and
$\ll, \gg$	bitwise shift left and right
$+, -$	unary plus and minus
$\sim$	bitwise not

Figure 3: Operations permitted in Computed syntax.

- The inherited *environment* attribute e, which is a map from variables to types on both the statement and expression non-terminals S and E.
- The synthesized *type* attribute t on expressions E, which contains the type of the expression.

With those, the SDD can be expressed as shown in Figure 1 using the helpers Extend, $\mathrm{Defined}$, and Lookup to build an extended environment with an additional type declaration, check for existence, and look one up, respectively, and Unif to find the type of an arithmetic operation with operands of two specific types.

We can translate this to the HACS in Figure 2, with the following schemes:

- Unif – unify two types as needed for typing the operators.
- Ee – scheme to propagate the e attribute over Exp values.
- Se – scheme to propagate the e attribute over Stat values.
- SeB – helper scheme to propagate the e attribute from left to right part of assignment statements.

The environment helpers get translated to native HACS environment patterns as follows:

- A "$\mathrm{Defined}(N.e, x)$" test is encoded by having two rules: one for the "true" branch with the constraint ↓e{⟦x⟧} in a pattern, and one for the "false" case with the constraint ↓e{⟦¬x⟧} in the pattern.
- "$\mathrm{Lookup}(N.e, x)$" is encoded by adding a constraint ↓e{⟦x⟧:#} in a pattern, which then binds the meta-variable # to the result of the lookup. (This will imply the "defined" pattern discussed above.)
- "$\mathrm{Extend}(N.e, x, V)$" is encoded by additing a constraint ↓e{⟦x⟧:V} in the *replacement*.

8 Compile-time Computations

We may sometimes need to compute helper values, most commonly for counting. HACS supports this through a special sort, Computed, which has special syntax for operations on primitive values.

The mechanism is quite simple: in rule replacements, arguments of sort Computed can contain expressions in ⟦⟧s, which include

- References to meta-variables of sort Computed.

- Integers in decimal and 0x-hexadecimal notation.
- Standard operators and parentheses.

8.1 Example (count)**.** Consider the list from Example 5.4. The following computes the length of the list.

```
sort Computed | scheme ListLength(List, Computed) ;
ListLength(⟦ ⟨Elem#1⟩ ⟨List#2⟩ ⟧, #n) →ListLength(#2, ⟦#n + 1 ⟧) ;
ListLength(⟦ ⟧, #n) →#n ;
```

9 Examples

Once the structure of a specification is clear, we can start analyzing and manipulating our internal representation. In this section we work through some examples of this.

9.1 Remark. Before we start, here is a short checklist of recommended practices for HACS processing:

- Separate token names from sort names, for example by using ALL CAPS for token names.
- If you use some kind of identifier then declare it as a special sort like

  ```
  sort Symbol | symbol ⟦⟨SYMBOL⟩⟧;
  ```

  with SYMBOL defined as a token that always allows numeric extensions, like

  ```
  token SYMBOL | [a−z]+ ('_' [0−9]+)* ;
  ```

  When you do this then you are allowed to use symbols of the specified kind in patterns, *i.e.*, a rule lile

  ```
  F(⟦x_1⟧) →⟦x_1 + x_2 ⟧;
  ```

  will allow F to be applied to *any* Symbol, with x_1 being the *representative* of that symbol for the purpose of the rule. Similarly, x_2 will correspond to a *globally fresh* symbol, as it only occurs left of the →.

9.2 Example (finding cats)**.** The small example in Figure 4 illustrates how to test for equality using a non-linear rule in line 12 combined with a "catch-all" default rule in line 13.

Note that we cannot use ⟦cat⟧ directly in a pattern: patterns are restricted to *syntactic cases* of the grammar. Also note that we have here defined the Boolean sort to have syntactic values rather than just constructors: this allows us to print them.

Here is a possible run using this command:

```
$ hacs IsCat.hx
...
$ ./IsCat.run --scheme=IsCat --term="dog"
False
$ ./IsCat.run --scheme=IsCat --term="cat"
True
```

```
module org.crsx.hacs.samples.IsCat {                                    1

token WORD | [A−Za−z]+ ;                                                3
main sort Word | ⟦⟨WORD⟩⟧ ;                                             4

sort Boolean | ⟦True⟧ | ⟦False⟧ ;                                       6

sort Boolean | scheme IsCat(Word) ;                                     8
IsCat(#word) →IsSameWord(#word, ⟦cat⟧) ;                                9

sort Boolean | scheme IsSameWord(Word, Word) ;                          11
IsSameWord(#, #) →⟦True⟧ ;                                              12
default IsSameWord(#1, #2) →⟦False⟧ ;                                   13
}                                                                       14
```

Figure 4: *IsCat.hx*: Finding cats.

9.3 Example (set of words)**.** One common task is to synthesize a set from some syntactic construct, and subsequently search the set. Figure 5 shows a small toy syntax allowing simple queries of word set membership.

The example uses some new mechanisms for synthesizing the set:

- A helper z synthetic attribute contains a *set* of word tokens, which is indicated by the attribute declaration ↑z{WORD} in line 9.
- We associate a z set with all values of the syntactic sort List in line 10.
- Lines 11 and 12 capture the synthesis of the set. Line 12 captures the simple case where a singleton list synthesizes a singleton set.
- Line 11 has a few more notations in play. First, the *pattern* part of the rule includes the inner pattern ↑z{:#ws}. This specifies that the special meta-variable ":#ws" captures all the existing members of the z set. Second, the result of the rule is to add *two* new things to the top level of the rule: ↑z{:ws} ↑z{#w}. This adds *both* the existing members (just matched) *and* the one new member #w to the result set.
- Lines 24 and 25 are almost the same: the one difference is that 24 matches sets that contain the #w word, whereas 25 matches sets that do not because of the ¬ logical negation sign.

We can run the example as follows:

```
energon1[~]$ hacs WordSet.hx
energon1[~]$ ./WordSet.run --scheme=Check --term="a in a,b,b,a"
Yes, the list contains a.
energon1[~]$ ./WordSet.run --scheme=Check --term="Foo in Bar"
No, the list does not contain Foo.
```

9.4 Example (map of words)**.** Figure 6 shows how a map can be synthesized and then used as an environment. The pattern is similar to the set example, except here we not only synthesize the map attribute m but also "copy" it over to an inherited map – an environment – e. Notice these extras:

```
module org.crsx.hacs.samples.WordSet {                                                    1

// Simple word membership query.                                                          3
main sort Query | ⟦ ⟨WORD⟩ in ⟨List⟩ ⟧;                                                   4
sort List | ⟦ ⟨WORD⟩, ⟨List⟩ ⟧| ⟦ ⟨WORD⟩ ⟧;                                              5
token WORD | [A−Za−z0−9]+ ;                                                              6

// Collect set of words.                                                                  8
attribute ↑z{WORD} ;                                                                      9
sort List | ↑z ;                                                                         10
⟦ ⟨WORD#w⟩, ⟨List#rest ↑z{:#ws}⟩ ⟧↑z{:#ws} ↑z{#w} ;                                      11
⟦ ⟨WORD#w⟩ ⟧↑z{#w} ;                                                                     12

// We'll provide the answer in clear text.                                                14
sort Answer                                                                               15
| ⟦Yes, the list contains ⟨WORD⟩.⟧                                                       16
| ⟦No, the list does not contain ⟨WORD⟩.⟧                                               17
;                                                                                         18

// Check is main query scheme, which gives an Answer.                                     20
sort Answer | scheme Check(Query) ;                                                       21

// The main program needs the synthesized list before it can check membership.            23
Check( ⟦⟨WORD#w⟩ in ⟨List#rest ↑z{#w}⟩ ⟧) →⟦Yes, the list contains ⟨WORD#w⟩.⟧ ;          24
Check( ⟦⟨WORD#w⟩ in ⟨List#rest ↑z{¬#w}⟩ ⟧) →⟦No, the list does not contain ⟨WORD#w⟩.⟧ ;  25
}                                                                                         26
```

Figure 5: *WordSet.hx*: Sets of Words.

- The map attribute is synthesized in lines 12–13 just like the set attribute was in the previous example, the only difference is that the map of course includes both a key and value.
- In line 23 we simply capture all the "mappings" of the m attribute with the special :#ms pattern, which we then *reuse* to populate the e environment.
- We actually combine the distribution of the inherited map with a recursive transformation that replaces words, in lines 26–34. The two rules for an initial WORD are mutually exclusive because the pattern in line 26 requires that the word is present with a mapping in the e attribute, whereas the pattern in line 31 requires that the word is not present.

Here is a run demonstrating the program:

```
energon1[~]$ hacs WordMap.hx
energon1[~]$ ./WordMap.run --scheme=Substitute --term="a:b in a b b a"
b b b b
energon1[~]$ ./WordSet.run --scheme=Substitute --term="k:v in a b c"
a b c
```

9.5 Example (word substitution)**.** Figure 7 shows a HACS program to collect substitutions from a document and apply them to the entire document. Notice the following:

```
module org.crsx.hacs.samples.WordMap {

// Simple word map over list.
main sort Query | ⟦ ⟨Map⟩ in ⟨List⟩ ⟧ ;
sort List | ⟦ ⟨WORD⟩ ⟨List⟩ ⟧| [[]] ;
sort Map | ⟦⟨WORD⟩ : ⟨WORD⟩ , ⟨Map⟩ ⟧| ⟦⟨WORD⟩ : ⟨WORD⟩ ⟧;
token WORD | [A−Za−z0−9]+ ;

// Collect word mapping.
attribute ↑m{WORD:WORD} ;
sort Map | ↑m ;
⟦ ⟨WORD#key⟩ : ⟨WORD#value⟩, ⟨Map#map ↑m{:#ms}⟩ ⟧↑m{:#ms} ↑m{#key:#value} ;
⟦ ⟨WORD#key⟩ : ⟨WORD#value⟩ ⟧↑m{#key:#value} ;

// Main program takes a Query and gives a List.
sort List | scheme Substitute(Query) ;

// Environment for mappings during List processing.
attribute ↓e{WORD:WORD} ;
sort List | scheme ListE(List) ↓e ;

// The main program needs the synthesized map before it can substitute.
Substitute( ⟦ ⟨Map#map ↑m{:#ms}⟩ in ⟨List#list⟩ ⟧) →ListE( #list ) ↓e{:#ms} ;

// Replace any mapped words.
ListE( ⟦ ⟨WORD#word⟩ ⟨List#words⟩ ⟧↑#syn ) ↓e{#word : #replacement}
→
⟦ ⟨WORD#replacement⟩ ⟨List ListE(#words)⟩ ⟧↑#syn
;

ListE( ⟦ ⟨WORD#word⟩ ⟨List#words⟩ ⟧↑#syn ) ↓e{¬#word}
→
⟦ ⟨WORD#word⟩ ⟨List ListE(#words)⟩ ⟧↑#syn
;

ListE( ⟦ ⟧ ↑#syn ) →⟦ ⟧ ↑#syn ;
}
```

Figure 6: *WordMap.hx*: Apply Word Substitution as Map.

```
module org.crsx.hacs.samples.WordSubst {

// Grammar.
sort Units | ⟦ ⟨Unit⟩ ⟨Units⟩ ⟧ | ⟦⟧ ;
sort Unit | ⟦⟨Variable⟩=⟨NAT⟩⟧ | ⟦⟨Variable⟩⟧ | ⟦⟨NAT⟩⟧ | ⟦{ ⟨Units⟩ } ⟧;
sort Variable | symbol ⟦⟨ID⟩⟧ ;

token ID | [A−Za−z]+ ;
token NAT | [0−9]+ ;
space [\ \t\n\r] ;

// Helper Subst structure: lists of variable-NAT pairs.
sort Subst | MoreSubst(Variable, NAT, Subst) | NoSubst ;

// Append operation for Subst structures.
| scheme SubstAppend(Subst, Subst) ;
SubstAppend(MoreSubst(#var, #nat, #subst1), #subst2) →MoreSubst(#var, #nat, SubstAppend(#subst1, #subst2)) ;
SubstAppend(NoSubst, #subst2) →#subst2 ;

// Attributes.
attribute ↑subst(Subst) ;          // collected Subst structure
attribute ↓env{Variable:NAT} ; // mappings to apply

// Top scheme.
main sort Units | scheme Run(Units) ;
Run(#units) →Run1(#units) ;

// Strategy: two passes.
// 1. force synthesis of subst attribute.
// 2. convert subst attribute to inherited environment (which forces replacement).

| scheme Run1(Units) ;
Run1(#units ↑subst(#subst)) →Run2(#units, #subst) ;

| scheme Run2(Units, Subst) ↓env ;
Run2(#units, MoreSubst(#var, #nat, #subst)) →Run2(#units, #subst) ↓env{#var : #nat} ;
Run2(#units, NoSubst) →Unitsenv(#units) ;

// Synthesis of subst.

sort Units | ↑subst ;
⟦ ⟨Unit #1 ↑subst(#subst1) ⟩⟨Units #2 ↑subst(#subst2)⟩ ⟧↑subst(SubstAppend(#subst1, #subst2)) ;
⟦ ⟧ ↑subst(NoSubst) ;

sort Unit | ↑subst ;
⟦v=⟨NAT#n⟩⟧ ↑subst(MoreSubst(⟦v⟧, #n, NoSubst)) ;
⟦v⟧ ↑subst(NoSubst) ;
⟦⟨NAT#n⟩⟧ ↑subst(NoSubst) ;
⟦ { ⟨Units#units ↑subst(#subst)⟩ } ⟧↑subst(#subst) ;

// Inheritance of env combined with substitution.

sort Units | scheme Unitsenv(Units) ↓env ;
Unitsenv( ⟦ ⟨Unit#1⟩ ⟨Units#2⟩ ⟧↑#s ) →⟦⟨Unit Unitenv(#1)⟩ ⟨Units Unitsenv(#2)⟩ ⟧↑#s ;
Unitsenv( ⟦⟧↑#s ) →⟦ ⟧↑#s ;

sort Unit | scheme Unitenv(Unit) ↓env ;
Unitenv( ⟦v=⟨NAT#n⟩ ⟧↑#s) →⟦v=⟨NAT#n⟩⟧↑#s ;
Unitenv( ⟦v⟧ ) ↓env{⟦v⟧:#n} →⟦⟨NAT#n⟩⟧ ;
Unitenv( ⟦v⟧↑#s ) ↓env{¬⟦v⟧} →⟦v⟧↑#s ;
Unitenv( ⟦⟨NAT#n⟩⟧↑#s ) →⟦⟨NAT#n⟩⟧↑#s ;
Unitenv( ⟦{ ⟨Units#units⟩ } ⟧↑#s ) →⟦{ ⟨Units Unitsenv(#units)⟩ } ⟧↑#s ;
}
```

Figure 7: *WordSubst.hx*: Combining list, maps, and transformation.

- The strategy is a typical two-pass one: first one pass to collect the substitutions into a synthesized attribute, then a second pass where the full list of substitutions is applied everywhere.
- We have chosen to synthesize the map as a *data structure* instead of a native HACS map as in the previous Example 9.4 because we here need to *append* two maps (in line 42), which is not supported for the native maps. The synthesis happens in lines 41–49.
- We translate the synthesized map in list form into a native HACS map before starting the second pass: notice how Run2 starts by recursing over the list of substitutions, inserting each into the carried inherited env map. Since the map is consumed from left to right, the *latest* substitution for any variable is always used.
- Since the inheritance schemes for env in lines 53–63 are doing a recursive traversal of the term, we benefit by building the actual substitutions into the traversal.
- In the inheritance rules we are careful to preserve the synthesized attributes only when the term does not change. In our case this is manifest by just the rule in line 59 not including the ↑#s marker to capture and copy the synthesized attributes; in general this should be considered for every situation.

Here is a run with this system:

```
energon1[~]$ hacs WordSubst.hx
...
energon1[~]$ ./WordSubst.run --scheme=Run --term="a=1 a"
 a=1  1
energon1[~]$ ./WordSubst.run --scheme=Run --term="b a {a=1 b=2}"
 2  1     a=1  b=2
energon1[~]$ ./WordSubst.run --scheme=Run --term="{a=1 b=2 c=3} a b c {a=4} a b c"
  {  a=1  b=2  c=3     }   4  2  3   {  a=4   }   4  2  3
```

The last example shows how the latest substitution for a "wins."

A Manual

This appendix is an evolving attempt at giving a systematic description on HACS.

A.1 Manual (grammar structure)**.** A HACS compiler is specified as a single *.hx* module file with the following structure:

```
module modulename
{
  Declarations
}
```

where the *modulename* should be a Java style fully qualified class name with the last component is capitalized, like `org.crsx.hacs.samples.First`. The individual sections specify the compiler, and the possible contents is documented in the manual blocks throughout this document.

A.2 Manual (lexical declarations)**.** A token is declared with the keyword token followed by the token (sort) name, a | (vertical bar), and a *regular expression*, which has one of the following forms (with increasing order of precedence):

1. Several alternative regular expressions can be combined with further | characters.

2. Concatenation denotes the regular expression recognizing concatenations of what matches the subexpressions.
3. A regular expression (of the forms following this one) can be followed by a *repetition marker*: ? for zero or one, + for one or more, and ∗ for zero or more.
4. A simple word without special characters stands for itself.
5. A string in single or double quotes stands for the contents of the string except that \ introduces an *escape code* that stands for the encoded character in the string (see next item).
6. A stand-alone \ followed by an *escape code* stands for that character: escape codes include the usual C and Java escapes: \n, \r, \a, \f, \t, octal escapes like \177, special character escapes like \\, \', \", and Unicode hexadecimal escapes like \u27e9.
7. A *character class* is given in [ ], with these rules:
   (a) if the first character is ^ then the character class is negated;
   (b) if the first (after ^) character is] then that character is (not) permitted;
   (c) a \ followed by an *escape code* is encountered then it stands for the encoded character;
   (d) two characters connected with a − (dash) stands for a single character in the indicated (inclusive) *range*.

   Note that a character class cannot be empty, however, [^] is permitted and stands for all characters.
8. The . (period) character stands for the character class [^\n].
9. A nested regular expression can be given in ( ).
10. An entire other token T can be included (by literal substitution, so recursion is not allowed) by writing ⟨T⟩ (the angle brackets are unicode characters U+27E8 and U+27E9). Tokens declared with token fragment can *only* be used this way.
11. The special declaration space defines what constitutes white space for the generated grammar. (Note that this does not influence what is considered space in the specification itself, even inside syntax productions.) A spacing declaration permits the special alternative nested declaration for nested comments, illustrated by the following, which defines usual C/Java style spacing with comments as used by HACS itself:

    space [\t\f\r\n] | nested "/*" "*/" | "//" .* ;

Notice that spacing is not significant in regular expressions, except (1) in character classes, (2) in literal strings, and (3) if escaped (as in \ ).

A.3 Manual (syntactic sorts)**.** Formally, HACS uses the following notations for specifying the syntax to use for terms.

1. HACS *production names* are capitalized words, so we can for example use Exp for the production of expressions. The name of a production also serves as the name of its *sort*, *i.e.*, the semantic category that is used internally for abstract syntax trees with that root production. If particular instances of a sort need to be referenced later they can be *disambiguated* with a #i suffix, *e.g.*, Exp#2, where i is an optional number or other simple word.
2. A sort is declared by one or more sort declarations of the name optionally followed by a number of *abstract syntax production* alternatives, each starting with a |. A sort declaration sets the *current sort* for subsequent declarations and in particular any stand-alone production alternatives. All sort declarations for a sort are cumulative.
3. Double square brackets ⟦. . . ⟧ (unicode U+27E6 and U+27E7) are used for *concrete syntax* but can contain nested angle brackets ⟨. . . ⟩ (unicode U+27E8 and U+27E9) with *production references* like ⟨Exp⟩ for an expression (as well as several other things that we will come to later). We for example write ⟦⟨Exp⟩+⟨Exp⟩⟧ to describe the form where two expressions are separated by a + sign.

4. Concrete syntax specification can include ¶ characters to indicate where *newlines* should be inserted in the printed output. (The system can also control indentation but that is not enabled yet.)
5. A trailing @p for some precedence integer p indicates that either the subexpression or the entire alternative (as appropriate) should be considered to have the indicated precedence, with higher numbers indicating higher precedence, *i.e.*, tighter association. (For details on the limitations of how the precedence and left recursion mechanisms are implemented, see Appendix C.)
6. sugar ⟦...⟧→... alternatives specify equivalent forms for existing syntax: anything matching the left alternative will be interpreted the same as the right one (which must have been previously defined); references must be disambiguated.
7. If a production contains only a reference to a token, where furthermore the token is defined such that it can end with _n (an underscore followed by a count), then the sort can be qualified as a symbol sort, which can be used for variables and binders.

A.4 Manual (parsed terms)**.** The term model includes *parsed terms*.

1. Double square brackets ⟦...⟧ (unicode U+27E6 and U+27E7) can be used for *concrete terms*, provided the *sort* is clear, either
   (a) by immediately prefixing with the sort (as in Exp⟦1+2⟧), or
   (b) by using as the argument of a defined constructor (as IsType(⟦mytype⟧)), or
   (c) by usng as an attribute value, or
   (d) by using as a top level rule pattern or replacement term with a defined current sort.
2. Concrete terms can contain nested raw terms in ⟨...⟩ (unicode U+27E8 and U+27E9). Such nested raw terms *must* have an explicit sort prefix.
3. The special term error⟦...⟧ will print the error message embedded in ⟦...⟧, where one is permitted to embed symbol-declared variables in ⟨...⟩.

A.5 Manual (raw terms, schemes, and rules)**.** "Raw" declarations consist of the following elements:

1. A *constructor* is a capitalized word (similar to a sort name but in a separate name space).
2. A *variable* is a lower case word (subject to scoping, described below).
3. A sort can be given a *semantic production* as a | (bar) followed by a *form*, which consists of a constructor name, optionally followed by a list of the subexpression sorts in parenthesis.
4. A semantic production can be qualified as a scheme, which marks the declared construction as a candidate for rewrite rules (defined below).
5. A *raw term* is either a *construction*, a *variable use*, or a *meta-application*, as follows
   (a) A *construction* term is a constructor name followed by an optional ()ed ,-separated list of subterms.
   (b) A *variable use* term is a variable, subject to the usual lexical scoping rules.
   (c) A *meta-application* term is a *meta-variable*, consisting of a # (hash) followed by a number or word and optionally by a meta-argument list of ,-separated terms enclosed in []. Examples include #t1 (with no arguments), #[a,b,c], and #1[OK,#].
6. A term can have a *sort prefix*. So the term Type Unif(Type #t1, Type Float) is the same as Unif(#t1,Float) provided Unif was declared with the raw production |Unif(Type,Type).
7. A *rewrite rule* is a pair of terms separatede by → (arrow, U+2192), with a few additional constraints: in the rule $p \to t$, p must be a *pattern*, which means it must be a construction term that has been declared as a scheme (syntactic or raw) and with the restriction that all contained arguments to meta-applications must be bound variables, and all meta-applications in t must have meta-variables that also occur in p with the same number of meta-arguments.

   Rule declarations must either occur with the appropriate current sort or have a pattern with a sort prefix.

8. One rule per scheme can be prefixed with the qualifier default. If so then the pattern can have no structure: all subterms of the pattern scheme construction must be plain meta-applications. Such a default rule is applied *after* it has been ensured that all other rules fail for the scheme.
9. Finally, a rule can be prefixed with the word rule for clarity.

Rules are used for *rewriting*, a definition of which is beyond the scope of this document; please refer to the literature on higher order rewriting for details [2].

A.6 Manual (attributes and synthesis rules)**.**

1. Attributes are declared by attribute declarations followed by an *attribute form* of one of the following shapes:
   (a) ↑Name(ValueSort) defines that the synthesized attribute Name has ValueSort values;
   (b) ↑Name{KeySort} defines that the synthesized attribute Name is a set of KeySort values;
   (c) ↑Name{KeySort:ValueSort} defines that the synthesized attribute Name is a map from KeySort to Valuesort values;
   (d) ↓Name(ValueSort), ↓Name{KeySort}, and ↓Name{SymbolSort:ValueSort} similarly for inherited attributes;
2. One can add a simple *synthesized attributes* after a raw data term as ↑*name*(*value*), where the *name* is an attribute name and the *value* can be any term.
3. Simple *inherited attributes* are added similarly after a raw scheme term as ↓*name*(*value*).
4. An *inherited symbol table attribute extension* is added to a raw scheme term as ↓*name*{*symbol*:*value*}, where the *symbol* is either a variable or a constant (of the appropriate sort).
5. A *synthesized attribute reference* has the simple form ↑*name*; and declares that the current sort synthesizes *name* attributes.
6. A scheme declaration can include *inherited attribute references* of the form ↓*name*, which declares that the scheme inherits the *name* attributes.
7. A *synthesis rule* is a special rule of the form $t \uparrow name(t')$, where the term t may contain subterms with attribute constraints. The rule specifies how terms of the current sort and shape t synthesize *name* attributes.
8. In *rules* one can use the special forms ↑#m, which captures *all* synthesized attribute values; ↑t{:#ms} (↓t{:#ms}), which captures the full set of keys or key-value mappings of the t synthesized (inherited) attribute.

Inherited attributes are managed with regular rules (for schemes) with inherited attribute constraints and extensions.

A.7 Manual (building and running)**.** To translate a HACS script to an executable, run the *hacs* command, which generates a number of files under a *build* subdirectory, as well as the main script with a *.run* extension. The script accepts a number of options:

1. `--sort=`*Sort* sets the expected sort (and thus parser productions) for the input to *Sort*. The input is read, normalized, and printed.
2. `--scheme=`*Constructor* sets the computation for the compiler to *Constructor*, which must be a unary raw scheme; the argument sort of *Constructor* defines the parser productions to use. The input is read, wrapped in the action, normalized, and printed.
3. `--term=`*text* use the *text* as the input.
4. `--input=`*file* (or just the *file*) reads the input from *file*.
5. `--output=`*file* sends the input to *file* (the default is the standard output).

6. `--errors` reports details of errors found by subprocesses.
7. `--verbose=`*n* sets the verbosity of the underlying CRSX rewrite engine to n. The default is 0 (quiet) but 1–3 are useful (above 3 you get a lot of low level diagnostic output).
8. `--parse-verbose` activates (very!) verbose output from JavaCC of the parsing.

You must provide one of `--sort` or `--scheme`, and one of `--term` and `--input`.

Notice that the *.run* script has absolute references to the files in the *build* directory, so the latter should be moved with care.

B Common Errors

In this appendix we list some of the more common of what can be called the "error messages" of HACS. *Note* that most of these only come out when HACS is run with the `-e` option.

B.1 Error (HACS syntax)**.**

```
Exception in thread "main" java.lang.RuntimeException: net.sf.crsx.CRSException:
 Encountered " "." ". "" at line 35, column 6.
Was expecting one of:
    <MT_Repeat> ...
    "%Repeat" ...
    <MT_Attributes> ...
```

This error message from the `hacs` command indicates a simple syntax errors in the *.hx* file.

B.2 Error (user syntax)**.**

```
Exception in thread "main" java.lang.RuntimeException:
  net.sf.crsx.CRSException: net.sf.crsx.parser.ParseException:
mycompiler.crs: Parse error in embedded myDecSome term at line 867, column 42:
 ⟦ $TA_Let2b ⟨Dec (#d)⟩{ ⟨DecSome (#ds)⟩} ⟧ at line 867, column 42
 Encountered " "\u27e9" "\u27e8Dec (#d)\u27e9 "" at line 867, column 53
...
```

This indicates a concrete syntax error in some parsed syntax—inside ⟦...⟧—in the *.hx* file. The offending fragment is given in double angles in the message. Check that it is correctly entered in the HACS specification in a way that corresponds to a syntax production. Note that the line/column numbers refer to the generated *build/...Rules.crs* file, which us not immediately helpful (this is a known bug). In error messages a sort is typically referenced as a lower case prefix followed by the sort name—here myDecSome indicates that the problem is with parsing the DecSome sort of the My parser.

B.3 Error (JavaCC noise)**.**

```
Java Compiler Compiler Version ??.??_?? (Parser Generator)
(type "javacc" with no arguments for help)
Reading from file FirstHx.jj . . .
Warning: Choice conflict involving two expansions at
         line 3030, column 34 and line 3033, column 8 respectively.
         A common prefix is: "{" <T_HX_VAR>
         Consider using a lookahead of 3 or more for earlier expansion.
Warning: Line 4680, Column 18: Non-ASCII characters used in regular expression.
Please make sure you use the correct Reader when you create the parser,
 one that can handle your character set.
File "TokenMgrError.java" does not exist.  Will create one.
File "ParseException.java" does not exist.  Will create one.
File "Token.java" does not exist.  Will create one.
```

```
File "SimpleCharStream.java" does not exist.  Will create one.
Parser generated with 0 errors and 1 warnings.
Note: net/sf/crsx/samples/gentle/FirstParser.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
```

These are "normal" messages from JavaCC. Yes, the choice conflict is annoying but is in fact safe.

B.4 Error (missing library)**.**

```
gcc -std=c99 -g    -c -o crsx_scan.o crsx_scan.c
crsx.c:11:30: fatal error: unicode/umachine.h: No such file or directory
```

The HACS tools only use one library in C: ICU. You should get the *libicu-dev* package (or similar) for your system.

B.5 Error (meta-variable mistake)**.**

```
Error in rule Tiger-Ty99_9148-1: contractum uses undefined meta-variable (#es)
Errors prevent normalization.
make: *** [pr3.crs-installed] Error 1
```

A rule uses the metavariable #es in the replacement without defining it in the corresponding pattern.

B.6 Error.

```
/home/krisrose/Desktop/teaching/.../hacs/cookmain PG pr3.hxt > pr3.pg
cookmain: crsx.c:528: bufferEnd: Assertion
   '(((childTerm)->descriptor == ((void *)0)) ? 0 :
        (childTerm)->descriptor->arity) == bufferTop(buffer)->index' failed.
/bin/sh: line 1: 14278 Aborted
  (core dumped) /home/krisrose/Desktop/teaching/.../hacs/cookmain PG pr3.hxt > pr3.pg
```

This indicates an arity error: a raw term in the *.hx* file does not have the right number of arguments.

B.7 Error.

```
// $Sortify
// $[Load, ".../build/edu/nyu/csci/cc/fall14/Pr2Solution.hx", "pr2solutionMeta_HxModule"]
Exception in thread "main" edu.nyu.csci.cc.fall14.TokenMgrError:
   Lexical error at line 184, column 31.  Encountered: "t" (116), after : "Call"
```

This indicates that you have an undefined symbol of sort error in the *.hx* file: the symbol starting with Callt is either undefined or used in a location where it does not match the required sort.

B.8 Error.

```
// $Sortify
// $[Load, ".../build/edu/nyu/csci/cc/fall14/Pr2Solution.hx", "pr2solutionMeta_HxModule"]
Exception in thread "main" java.lang.RuntimeException: net.sf.crsx.CRSException:
   Encountered " ")" ") "" at line 255, column 112.
Was expecting one of:
    "," ...
```

This indicates that you have an incorrect number of arguments in the *.hx* file: here insufficient arguments (encountering a) instead of ,); a similar but opposite error is given when excess arguments are present.

B.9 Error.

```
/home/krisrose/Desktop/teaching/.../hacs/cookmain PG pr3.hxt > pr3.pg
cookmain: crsx.c:528: bufferEnd: Assertion
   '(((childTerm)->descriptor == ((void *)0)) ? 0 :
        (childTerm)->descriptor->arity) == bufferTop(buffer)->index' failed.
/bin/sh: line 1: 14278 Aborted
  (core dumped) /home/krisrose/Desktop/teaching/.../hacs/cookmain PG pr3.hxt > pr3.pg
```

This indicates an arity error: a raw term in the *.hx* file does not have the right number of arguments.

B.10 Error.

```
« $Print-Check[
...
»
```

This error from your *.run* script indicates that you requested a `--scheme` Check, which is not in fact declared as a scheme in the *.hx* file.

C Limitations

- At most one nested declaration per token.
- Precedence can only be used on self references, *i.e.*, ⟨E@2⟩ can only occur inside productions for the sort E.
- It is not possible to use binders and left recursion in the same production with the same precedence.
- Only *immediate* left recursion is currently supported, *i.e.*, the left recursion should be within a single production.
- Productions can share a prefix but only within productions for the same sort, and the prefix has to be literally identical unit by unit (except for left recursive precedence markers), *i.e.*,

  ```
  sort S | ⟦ ⟨A⟩ then ⟨B⟩ then C ⟧
         | ⟦ ⟨A⟩ then ⟨B⟩ or else D ⟧ ;
  ```

  is fine but

  ```
  sort S | ⟦ ⟨A⟩ then ⟨B⟩ then C ⟧
         | ⟦ ⟨A⟩ ⟨ThenB⟩ or else D ⟧ ;
  sort ThenB | ⟦ then ⟨B⟩ ⟧;
  ```

  is not.
- It is not possible to left-factor a binder (so multiple binding constructs cannot have the same binder prefix).
- Variables embedded in error⟦...⟧ instructions must start with a lower case letter.
- When using the symbol qualifier on a reference to a token then the token *must* allow ending in $\_n$ for n any natural number.
- When using the same name for a symbol inside of ⟦...⟧ and the corresponding raw variable outside of the ⟦⟧, then the common symbol and variable name must be a plain word starting with a lower case letter.
- Special terms like error⟦...⟧ cannot be used as raw subterms.
- The default rule qualifier is rather fragile and does not yet always work.

References

[1] Alfred V. Aho, Monica S. Lam, Ravi Sethi, , and Jeffrey D. Ullman. *Compilers: Principles, Techniques and Tools*. Pearson Education, Inc, 2006.

[2] Jan Willem Klop, Vincent van Oostrom, and Femke van Raamsdonk. Combinatory reduction systems: Introduction and survey. *Theor. Computer Science*, 121:279–308, 1993. doi:10.1016/0304-3975(93)90091-7.

[3] Donald E. Knuth. Semantics of context-free languages. *Mathematical Systems Theory*, 2(2):127–145, 1968. doi:10.1007/BF01692511.

[4] P. Naur et al. Report on the algorithmic language ALGOL 60. *Communications of the ACM*, 3:299–314, 1960.

[5] Kristoffer Rose. Combinatory reduction systems with extensions. SourceForge project http://crsx.sf.net, October 2012.

[6] Sreeni Viswanadha, Sriram Sankar, et al. *Java Compiler Compiler (JavaCC) - The Java Parser Generator*. Sun, 4.0 edition, January 2006. URL: https://javacc.dev.java.net/.

[7] Sreeni Viswanadha, Sriram Sankar, et al. *ICU – International Components for Unicode*. ICU Project Management Committee, 54 edition, October 2014. URL: http://site.icu-project.org/home.