ANTLR Yggdrasil Manual

Loring Craymer

0.5b2 Draft

Introduction

Welcome to ANTLR Yggdrasil! ANTLR Yggdrasil is a variant/extension of ANTLR, a powerful language translation package developed by Terence Parr. ANTLR Yggdrasil differs from the baseline ANTLR 3 in terms of both capabilities and philosophy of language translation. With ANTLR Yggdrasil, you should never need to depend on the capabilities of the target language when developing a translator; instead, the needed capabilities are directly supported by Yggdrasil syntax.

Yggdrasil is about syntax trees, and it implements a concept that I call "Tree Attribute Grammars". [A tree attribute grammar is an L-attributed grammar with grammar-level attributes which include an output tree and which specifies a single pass translation. Attribute synthesis rules are not supported per se; instead, semantics support is through a minimalist attribute algebra and templates.] Yggdrasil focuses on syntax trees, their decoration and construction. Thus the name "Yggdrasil": in Norse mythology, Yggdrasil is the World Tree that connects the nine realms of the Norse cosmos.

Philosophically, Yggdrasil is driven by the following:

1. The goal is rapid development of highly capable language processors.
2. Syntax trees are fundamental, and tree restructuring is the basic mechanism for language translation. The ^ and ! annotation developed by Parr for SORCERER is a powerful base for such restructuring, but the idea should be taken further. No restructuring code should occur in actions defined in the target language.
3. Users should not have to manually derive tree walker grammars. That is something that can be done automatically.
4. Translators be target language independent. It should take minimal effort to re-target a language translator to another implementation language.
5. Attributes should be first class syntactic elements of ANTLR. Attribute manipulation, including the retyping of tokens or syntax tree nodes and the replacement of token text, should be directly expressible in ANTLR/Yggdrasil syntax.
6. Code for attribute classes and other datatypes used in an Yggdrasil application should be generated by Yggdrasil whenever possible.
7. While target language actions are to be avoided during the process of translating language input, the translator may have to be integrated with code that processes the data once it is in that internal form. Not all translation applications need back-end processing, but integration with a back-end processor written in the target language should be painless.

Language Composition

ANTLR/Yggdrasil is composed of multiple domain specific languages. These are

1. An EBNF dialect for character recognition and token identification.
2. An EBNF dialect for parsing.
3. An EBNF dialect for tree walking.
4. Yggdrasil—a language for defining and manipulating attributes for constructing syntax trees

and graphs.

5. String Template—a language for processing attributes into text.

Language composition is not new--lex/yacc probably provides the seminal example—but the use of composition as a pattern for development seems not to have been previously formalized. ANTLR Yggdrasil not only shows the power of language composition—as a result of composition, it cleanly supports very rapid development of language processors—but it also provides a framework for composing grammars.

Yggdrasil is a domain specific language for attribute definition and manipulation. It implements an attribute algebra that is conceptually simple, but capable of generating very complex data structures for back-end processing. It is implemented as an island grammar integrated with the EBNF dialects from ANTLR 2.

String Template is an output language. Despite a deceptively simple syntax, it provides a language for solving almost any text generation problem imaginable. At this point, String Template has a standalone grammar; this is likely to continue to be the case as part of the String Template language is "any sequence of characters except an unescaped '$' or '<'": these represent data not processed by the StringTemplate runtime library.

ANTLR also integrates "pluggable" support for target languages—Java, C#, C/C++, etc.--to gain access to processing capabilities beyond those supplied by the five DSLs.

Attributes in Yggdrasil

Attributes are characteristics of objects. For our purposes, attributes are either represented by data items or "fields" of the object or are computed from fields in the object. In Yggdrasil, attributes may be associated with 1.) grammars, 2.) tokens, and 3.) AST nodes. Attributes may be hierarchical: compound attributes may themselves have attributes. Attribute names must begin with a lower case letter.

Attributes are translated to target language objects, and Yggdrasil attributes are strongly typed. Yggdrasil has syntax to directly support manipulation of attributes; except for the ^ and ! rewrites, tree rewrites are accomplished by assigning tokens or AST nodes to grammar attributes and later referencing them to instantiate them as nodes in the generated syntax tree.

Conceptually, the grammar-level attributes comprise a blackboard which can be written to or read from. There is dynamic scoping support for rules; that is, the current value of a grammar-level attribute can be saved on entry to a grammar rule and set to another value—including a "blank" attribute of the same type.

Attribute Typing

One unusual feature of Yggdrasil is that attributes are strongly typed. That is, attribute data types are defined as in the example below:

@{

```
        native atomic int;
        native atomic String;
        Payload {
                int type;
                String text;
        }
}
```

Yggdrasil then knows that an attribute of type "Payload" has attributes "type" of type int and "text" of type String. The "native" keyword specifies that this type is provided by the target language; an atomic type is primitive. Yggdrasil automatically generates code for defined types (excluding those declared "native"); however, the generated code may not compile: package and import information is not generated and must be supplied by the developer[1]. This example is contrived, in the sense that both int and String are built in primitives that are automatically translated by Yggdrasil into their target equivalent: in some possible targets, "int" would become an Integer object.

Yggdrasil built-in types consist of

1. atomic types
    1. String
    2. Integer
    3. TokenType – Integer alias for token types.
2. Compound types
    1. Payload (type, text)
    2. Carrier (down, right, type)
    3. Queue
    4. Table
    5. Text (StringTemplate). See "Action Templates" section, below.
    6. Object (everything is an Object, but when referenced, it is used as a catchall base for compound type definition)

Additionally, literals and user-defined atomic types are supported. Users may subclass any of the compound types. Attribute support for lexers is currently weak; it may be necessary to add character queues and possibly other character types.

Generics

To maintain strong typing, both Queue and Table are treated as generics. That is, a Queue data item is declared as

1 This may change, but for the moment I am opting for simplicity of implementation.

Queue<Type> a;

and Tables are declared as

Table <IndexType, DataType> t;

with Type, IndexType, and DataType replaced by the desired types. Currently, generic syntax is only supported for Queues and Tables; user-defined generics will be supported in later versions.

Literals

It is necessary to provide token type and string literals to support modification of tree nodes. Since attribute names must begin with a lower case letter (simplifies translation during code generation), tokens beginning with an '@' sign but not a lower case letter are taken as literal values. Thus

@NUMBER is the literal N UMBER, @123 is the literal 123, @"String" is "String", and @'a' is 'a'. Text literals might be altered in the translation process—by adding a prefix or suffix, for example—but are not translated into method calls. Numeric literals follow the Java conventions and also may be translated, but strings and character literals are not.

To support references to fields beginning with lower case letters, literal text may be prefixed with a '\'; thus @\bVar becomes bVar.

Note: No type checking is performed on literals by Yggdrasil; typographical mistakes are caught by either the target language compiler or at runtime.

The Yggdrasil Attribute Algebra

Yggdrasil supports four operations on attributes. Attributes may be

1. Assigned.
2. Instantiated in the syntax tree.
3. Constructed from other attributes; this includes construction of syntax tree fragments.

Attributes may also be computed; however, there is no special syntax for computed attributes. The computation simply occurs when an attribute is assigned to or read from.

This simple set of operations is sufficient to provide complete rewrite functionality. The syntax for these operations can be mingled with the ^ and ! annotation taken from SORCERER and ANTLR 2 and is intended to be intuitive and easily used. [Cases where it is not should be reported to the author so that they can be fixed.]

Assignment syntax is what might be expected:

a = B

Here, B is a token in the token stream to be recognized. It is then assigned to attribute a (a grammar-level attribute) and instantiated[2] in the output tree.

2 I experimented with an @a = B syntax for assignment without instantiation, but this becomes confusing when dealing with attribute to attribute assignments.

a.b.c = D

Again, D is a token to be recognized. It is assigned to the c "field" of the attribute occupying the b field of grammar-level attribute a. (In other words, it behaves just like the equivalent Java assignment statement.)

One of the items that was never taken off of the ANTLR 2 "to do" list was subrule labeling. The idea was that a labeled subrule such as sub:(A B^ C)* had a restricted scope for the ^ operation. Yggdrasil achieves this via assignment of a subrule:

sub = (A B^ C)*

assigns the tree generated from the subrule to the sub attribute. This is logically consistent: by assigning the subrule to an attribute, we can remove it from the output stream without side effects on other tokens matched within the rule by using the "!" suffix:

sub = (A B^ C)*!

Instantiation occurs by referencing the attribute with an @ prefix:

@a

instantiates attribute a. A typical use of assignment and instantiation to rearrange a syntax tree is

a = B!

C D

@a

This results in (C D B) in the output AST.

An alternative view of assignment and instantiation is that attributes can either be used as labels (a = A) or as input/output variables: @a = A can be interpreted as "A is input to a", and @a can be interpreted as "A is output from a into the tree undergoing construction).

Carrier variables can either be used to label a point in the constructed tree or to hold a subtree that has been constructed but not yet included in the output tree [either a = rule! Or a = (...)! -- the "!" suffix after an assignment constructs but does not include a subtree in the output tree.] Replication of subtrees is necessary to instantiate more than one copy in the output tree; that is done via the "copy" attribute. Note that the copies should be instantiated in the output tree before the original—otherwise, the original becomes a pointer into the tree and a copy may be larger than intended!

Construction of attributes is either coupled with an assignment or with instantiation in the output syntax tree. Attributes are constructed with an @(...) syntax. To construct a Payload in the output tree, for example, we might do

a = A

b = B

@(a.type b.text)

to construct a Payload that is of the same type as A, but has the text from B. The Payload would be of

default type for the grammar.

Alternatively, we can couple construction with assignment to take advantage of attribute typing:

```
a = A!
b = B!
c = @(a.type, B)`  @c
```

If attribute c has been declared as a MessyPayload, this would result in a call to

```
new MessyPayload(A.type, B)          // Java or C++
```

for construction since anything assigned to c must be a MessyPayload.

Note that the constructed attribute must be explicitly instantiated in the output tree (@c). There is an exception to this:

```
c = @( a.type, B)^
```

would instantiate the constructed token as the root of the current rule.

Constructed attributes need not be Payloads. Any attribute type may be constructed and assigned. However, the developer will need to define the appropriate constructors in the target language as part of the definition of an attribute type.

Special features

Within a rule, the current tree fragment (rule or assigned subrule) is referenced by "@$". You can assign @$, kill it (@$!), copy it (@$.copy) or reference fields in the current root or start node (this when the rule is constructing a sequence of nodes and does not have a root yet assigned).

Tree Construction

Tree construction in Yggdrasil follows the SORCERER and ANTLR 2 approach of using ^ and ! for structuring. I experimented with a tree construction syntax in ANTLR 2.8, but the attribute algebra of Yggdrasil makes anything beyond the ^ annotation unnecessary. To instantiate attribute a as the root of the current tree, simply reference @a^.

One feature that I have kept from ANTLR 2.8 is the idea of predicated tree construction. One of the objectives in designing tree structures is to remove semantic ambiguities, and so semantic differences are converted to syntactic differences in tree structure. Thus we might differentiate a keyword reference from a variable reference by constructing different trees

```
a = A
b = B
c = C
@{     ?{ vars.contains(a) }  @a  @b  @c
       |
```

```
                @a^  @b  @c
        }
```

to decide whether to construct A B C or #(A B C).

Construction predicates

One of the difficulties in using semantic predicates in conjunction with backtracking is that semantic context might change dynamically. Yggdrasil "solves" this problem by distinguishing between the use of semantic predicates for recognition and for construction. During recognition, attributes are not altered. Instead, attribute modification occurs during tree construction at a time when there can be no backtracking and changes in semantic context are deterministic. So for languages with truly ugly ambiguities—like C++--the general translation approach is to capture semantic differences in structuring the syntax tree and defer recognition until the next tree walker pass[3].

Target language translation of the attribute algebra

In general, attribute references are translated into "get/set" method calls. In Java, at least, neither class nor attribute names are mangled; this may change for other languages. C "class" names will likely require mangling due to the lack of package/namespace support.

Code generation and adaptation

Yggdrasil can generate code for all attribute classes. In fact, this was one of the primary motives for having strongly typed attributes: users should not have to create support code by hand that can be created by machine. The generated code is fairly simple—a generated class contains fields, get/set methods, and skeletal query functions that always return "false".

Most generated classes will be "good enough", but many will need to be customized. Areas of customization include

1. Recoding of query functions to provide utility.
2. Addition of set<attribute>(CustomPayload) methods.
3. Addition of computed get/set methods where desirable.

Regeneration of attribute type classes is a normal part of development, but care should be taken to avoid overwriting customized code. I find diff3 very useful for this. Invoking (GNU) diff3 -m <myfile> <base file> <new file> > tempfile; mv tmp <myfile> is a convenient way of updating the customized source file. It is probably good practice to generate files into a temporary directory, customize them, move them to a target directory, and then having the ant scripts or makefiles handle the update process.

Grammar attributes that are declared "public" will have globally visible get/set accessors. Sometimes,

3 Yggdrasil does not have a mechanism for detecting "improper" use of semantic predicates, and the developer is free to violate this principle. However, there is no guarantee that the generated translator will correctly recognize dynamic context.

attributes from one pass will need to be carried over to the next—symbol tables are good candidates—and public accessors make this possible.

Action Templates

A major goal of Yggdrasil has been to remove actions from tree construction to achieve target language independence. Having a syntactic interface to StringTemplate removes the need to generate target code through actions—a very important step. It is still not enough. For a large class of applications, the purpose of language processing is to build internal data structures, data structures which are then operated on by the application. For these, some form of actions seem necessary.

One way of achieving target independence is to implement a procedural language within ANTLR. That has never been a particularly attractive solution; one obvious side-effect would be to dramatically increase the size of the ANTLR runtime code. That would be a severe hindrance to porting the ANTLR runtime to a wide variety of target languages.

To avoid this, Yggdrasil is adopting a template approach to actions. Instead of including a behavioral syntax in Yggdrasil, actions are simply templates, templates which can reference both externally defined templates and grammar attributes. Target language code is then relegated to template files, template files which can be recoded for different targets.

The syntax for templates is

// ACTION or SEMPRED equivalent

action : “{“ (QUOTED_STRING | ID) (“:” (assignment)+)? (“}” | “}?”) ;

assignment : ID “=” attribute_expression “,” ;

QUOTED_STRINGs are anonymous templates, ID is a named template within the StringTemplateGroup assigned to UserLibrary (grammar option).

Template assignment

Action templates are also used for assignment to “Text” variables. In this case, the actions are not instantiated in the generated code; instead, the text generated by the templates is assigned. The result is a simple syntax for using templates for code generation.

Discussion

Action templates effect Model/View/Controller separation: a properly templated grammar can be developed for one target environment and ported to another by simply editing template files. For anyone who suffered through debugging ANTLR 2 C++ code (Ric Klaren did an amazing job providing good C++ support, but C++ debugging was exceedingly unpleasant because the debuggers found C++ templates remarkably opaque and passed that opacity back to the often confused developer), this is a ray of sunshine. Applications can be developed for a friendly target language like Java and converted to a less friendly target by replacing template files.

Unfortunately, this forces the early incarnations of Yggdrasil to depend on the StringTemplate library.

This increases the effort of targeting ANTLR and Yggdrasil to new languages, and adds to the amount of code packaged with the generated application. In later incarnations, I expect templates to be compiled into classes with the templates converted to arrays of objects (toString() then just iterates though the array, calling toString() for each of the objects in the array) and get/set methods generated for filling in the "holes". That should improve memory use, allow applications to be shipped without including sets of proprietary template files, and provide some (probably slight) speed improvements.

Lexical annotation

Yggdrasil supports the use of "!" in lexer rules without sacrificing performance in most cases. For strings in which "!" is used to remove leading and trailing characters, the begin and end pointers are adjusted accordingly. For strings in which internal characters are removed, a "copy on write" approach is taken: when an incipient gap is detected, characters are copied to an internal "copy" array, and the start and end positions of text in the token reference the copy array.

I intend to provide string and character attribute support in later versions; the ANTLR 2 lexer grammars were very powerful in part because of string editing support and the inclusion of attribute support should make it "easy" to write smart editors (the engine, not the display).

Keywords and strings

In some languages, keywords (and symbols) are reserved and can appear in no other identifier. In others, keywords are special only in context: in a statement, for example, "if" might occur as a keyword -- `if (foo) bar;` -- or as a variable: `int if = 0;`. Many of the challenging translation problems deal with "languages" which do not have reserved keywords. ANTLR 3's DFA-based recognition of reserved strings represents a huge performance improvement over the ANTLR 2 literal table, but only addresses the problem of reserved keywords.

Languages with non-reserved keywords were a problem for ANTLR 2 because of the lack of the semantic predicate hoisting that was introduced in PCCTS (ANTLR 1). ANTLR 3 has brought back predicate hoisting, so that it is once again possible to do

```
{LT(1).getText().equals("foo") }? "foo".
```

This, however, does not go far enough. Yggdrasil supports two kinds of "strings": keywords, including characters in the lexer, are distinguished by single quotes as in the baseline ANTLR 3. Doubly quoted strings, on the other hand, are matched by textual content" "foo" is converted to

```
{LT(1).getText().equals("foo") }? .
```

to provide context-dependent keyword matching.

Double-quoted strings in the lexer are translated to a sequence of characters. (The baseline ANTLR 3 uses singly-quoted strings for this. Yggdrasil does not support this use of singly quoted strings; instead, the model is that the keywords for the lexer are individual characters.

Multi-token lexing

Yggdrasil supports multi-token lexing. "protected" rules produce tokens, just as the default "public"

rules do, but they are not included in the master case statement (or case statements, when lexical modes are supported). For example, consider

```
FOO
        :
        'A'  'B'  'C'  BAR 'g' 'h'
        ;
protected
BAR :  'D' 'E'  ;
```

BAR tokens are only recognized within the context of FOO; that is, the sequence of characters ABCDEgh is recognized as

FOO BAR FOO

while an isolated occurrence of DE in the input stream is an error (unless handled elsewhere in the grammar).

Multi-token lexing will be supported at about the same time as template compilation, largely because I see using it in constructing the template recognizer within ANTLR Yggdrasil. The lexical attribute support will include support for sub-rule tokens and the construction of tokens with rearranged text: that is a natural analogue of "labeled sub-rule" isolation of ^ context.

Development Plan

0.5: First release.

0.6: Initial tree grammar generation. May not support all "^ in loop" cases.

0.7: "Yggdrasil in Yggdrasil". Removes dependency on ANTLR 2.7 runtime; should include lexer attribute and editing support. Most of this development is expected to go towards writing an ANTLR 2.7.6 to Yggdrasil translator. StringTemplate type support will be complete in this release.

0.8: Adaptation to support generation of C code. Supporting languages that support classes and objects should be easy. For C, however, the translation of attribute types will be a slight challenge.

0.9: Documentation and cleanup. Will probably include "compiled" StringTemplates.

1.0: Production release.

Appendix A: Syntax

Syntactic element	Meaning
(...)	Subrule
*	Suffix denoting zero or more instances
+	Suffix denoting one or more instances
?	Suffix denoting zero or one instances
Lower case name	Parser rule or attribute variable
Capitalized name	Token rule
'...'	Keyword (including lexer characters) with unique token type
"..."	String—either a context-dependent keyword (parser, tree parser) or a sequence of characters (lexer)
{ ... }	Action template
{ ... }?	Semantic predicate template
(...)=>	Syntactic predicate
|	Alternative separator ("OR")
..	Range operator – 'a' .. 'b'
~	Not operator ("anything but")
.	Wildcard
@	Attribute (read) or literal prefix
=	Attribute assignment (also used for token and option statements)
!	"Do not output" suffix
^	Token suffix—make this the root of the current subtree
:	Rule start
;	Rule end (has other uses)
grammar	
options { ... }	Options section
tokens { ... }	Tokens section: Includes "imaginary" tokens, named keywords, and specifies Payload types.
@{ ... }	Attribute definitions section at the grammar level; also used to surround predicated tree construction alternatives.
@(...)	Attribute constructor.

<...>	Suffix used to specify types (token or generic specification)

Appendix B: Options

Option	Argument use or meaning if =true.
ASTLabelType	
backtrack	Support syntactic predicates.
buildAST	Build a syntax tree for the grammar (Yggdrasil)
filter	
factory	Specifies Payload Factory class used in generated recognizer.
greedy	(rule and subrule only)
k	Lookahead value
language	Target language; Java if unspecified.
memoize	Keep information to avoid repeated parses during backtracking. (Not yet supported in Yggdrasil)
outputVocab	Assigns name to generated .tokens file (Yggdrasil only)
superclass	Name of class which recognizer extends.
TokenLabelType	
TokenVocab	Input .tokens file
userTemplates	Colon separated list of template groups used in actions (Yggdrasil only).

Appendix C: command line arguments

Argument	**meaning**
<filename>	Grammar file to be processed.
-o outputDir	specifies output directory
-lib dir	Specifies location of .tokens files
-userLib templateDir	Toplevel directory for user templates. Language-specific subdirectories are below this. (Yggdrasil only)
-gen dir	Directory for generated type classes.
-verbose	
-nfa	generate an NFA for each rule
-dfa	generate a DFA for each decision point
-debug	Generate an instrumented recognizer (Yggdrasil)
-trace	Generate a traced recognizer.
-report	print out a report about the grammar processed
-profile	Same as -debug (Yggdrasil)
-print	Print the processed grammar
Internal options (-X)	
-X	Prints help for X commands
-Xgrtree	Print the grammar tree
-Xdfa	Print the dfa as text
-Xnoprune	Do not test ebnf exits
-Xnocollapse	Collapse incident edges into dfa states
-Xdbgconversion	Dump debugging information during NFA to DFA conversion.
-Xmultithreaded	Run the analysis in two threads
-Xnomergestopstates	Do not merge stop states
-Xdfaverbose	Generate DFA states in DOT with NFA configs
-Xwatchconversion	Print a message for each NFA before converting
-XdbgST	put tags at start/stop of all templates in output

-Xm n	Sets recursion limit to n (max number of rule invocations)
-Xmaxdfaedges n	max "comfortable" number of edges for single DFA state
-Xconversiontimeout nsecs	set NFA conversion timeout for each decision

The ANTLR Yggdrasil main is in org.antlr_yggdrasil.Tool.