Declarative Specification of Template-Based Textual Editors

Tobi Vollebregt
tobivollebregt@gmail.com

Lennart C. L. Kats
l.c.l.kats@tudelft.nl

Eelco Visser
visser@acm.org

Software Engineering Research Group
Delft University of Technology
The Netherlands

ABSTRACT
Syntax discoverability has been a crucial advantage of structure editors for new users of a language. Despite this advantage, structure editors have not been widely adopted. Based on immediate parsing and analyses, modern textual code editors are also increasingly syntax-aware: structure and textual editors are converging into a new editing paradigm that combines text and templates. Current text-based language workbenches require redundant specification of the ingredients for a template-based editor, which is detrimental to the quality of syntactic completion, as consistency and completeness of the definition cannot be guaranteed.

In this paper we describe the design and implementation of a specification language for syntax definition based on templates. It unifies the specification of parsers, unparsers and template-based editors. We evaluate the template language by application to two domain-specific languages used for tax benefits and mobile applications.

Categories and Subject Descriptors
D.2.3 [Software Engineering]: Coding Tools and Techniques—pretty printers, program editors; D.3.1 [Programming Languages]: Procesors—parsing, code generation; D.3.1 [Programming Languages]: Formal Definitions and Theory—syntax; D.2.6 [Software Engineering]: Programming Environments; D.2.11 [Software Architectures]: Languages

1. INTRODUCTION
Language-aware structure editors provide a template-based paradigm for editing programs. They allow composing programs by selecting a template and filling the placeholders, which can again be extended using templates. A crucial advantage of structure editors is syntax discoverability, helping new users to learn a language by presenting possible syntactic completions in a menu. Structure editors can be automatically generated from a syntax definition. Notable projects aiming at automatic generation of structure editors include MPS [23] and the Intentional Domain Workbench [19]. Structure editors can be used for general-purpose languages or for domain-specific languages (DSLs). A modern example of the former is the structure editor in MPS [23], for extensible languages based on Java. An example of the latter category is the DSL for modeling tax-benefit rules developed by IT services company Capgemini using the Cheetah system. Cheetah’s facilities for discoverability and the use of templates are particularly effective to aid a small audience of domain expert programmers manage the verbose syntax based on legal texts.

Despite their support for discoverability, structure editors have not been widely adopted. Pure structure editors tend to introduce an increased learning curve for basic editing operations. For example, they only support copy-pasting operations that maintain well-formedness of the tree and require small, yet non-trivial “refactoring” operations for editing existing code, e.g. when converting an if statement to an if-else statement. They also lack integration with other tools and expose the user to vendor lock-in. Transferring code across tools requires a shared representation that is generally not available. With software engineering tools such as issue trackers, forums, search, and version control being based on text, a textual representation is preferable, but requires the use of a parser and a parseable language syntax. This forces tools based on structure editors to find new solutions to problems long solved in the text domain.

To alleviate the problems of structure editors, there has been a long history of hybrid structure editors that introduce textual editing features to structure editors [24, 18, 11]. Conversely, modern textual code editors such as those in Eclipse and Visual Studio are increasingly syntax-aware, based on parsers that run while a program is edited. Over time, they have acquired features ranging from code folding to syntactic completions allowing programmers to fill in textual templates. Indeed, structure and textual editors are converging into a new editing paradigm that combines text and templates.

In order to provide the advantages of text editing to the tax-benefit DSL of Capgemini, we converted the language from the Cheetah system, which uses a structure editor, to the parser-based Spoofax language workbench [10]. We quickly realized that it would be impossible for a user to write new models in such a verbose language in the textual editor of Spoofax, without accurate and complete syntax discovery. In syntax-aware text editors this discovery is provided in the form of syntactic completion. Accurate and complete syntactic completion depends critically on two features of a language workbench: first, the syntactic completion proposals presented to the user must be relevant and complete, and second, it must be feasible to create and maintain the specification necessary to make the editor aware of these completion proposals.
Text-based language workbenches currently require redundant specification of the ingredients for a template-based editor, i.e. concrete syntax, abstract syntax, completion templates, and pretty-print rules, which is detrimental to the quality of syntactic completion in syntax-aware editors. Evolution of the language requires maintenance of all ingredients in order to maintain completeness and consistency. It is tedious and therefore easy to make mistakes while adding or adapting a completion template for each new or modified language construct.

In this paper, we present the design of a template-based syntax definition language\(^1\) that unifies the specification of parsers, unparsers, and template-based editors in order to support the efficient construction of template-based editing facilities in textual editors. In order to improve the runtime support for these template-based editing facilities, we describe an approach to compute the set of applicable templates at the location of the cursor.

We have implemented the template language in an extension of the Spoofax language workbench [10] and validated the approach by applying the techniques in two mature DSLs.

We proceed as follows. In the next section, we provide background on language and editor implementation. In Section 3 we describe the design of a template-based specification language for syntax. We then describe how to generate template-based editors from such a language in Section 4 and how template-based editing can be supported in Section 5. In Section 6 we evaluate the approach. We discuss related work in Section 7 and present our conclusion in Section 8.

2. BACKGROUND

Modern integrated development environments (IDEs) contribute significantly to the productivity of software developers and the adoption of new languages. The development of a complete IDE from scratch is a significant undertaking. As an alternative, language workbenches [6, 10, 23] can generate a complete IDE plug-in and a compiler from high-level language definitions.

Key to the generation of full-featured editors is the use of an abstract representation of programs that is maintained as programs are edited. The abstract representation is used for editor services, i.e. facilities such as an outline of the program and reference resolving to navigate to the definition sites of identifiers. Some of these facilities can be implemented directly at the level of concrete syntax, using regular expressions or other forms of pattern matching, but a structured, abstract representation ensures a uniform interface for editor services.

A central part of the definition of a language is the mapping between the concrete syntax of a language and its abstract representation. In structure editors, the mapping is defined as a projection from abstract representation to concrete syntax. In textual editors, the mapping is defined through a parser that constructs the abstract representation from a textual concrete syntax. Syntax-aware textual editors also apply a reverse mapping in editor services such as content completion, pretty printing, code formatting, and refactoring. In particular, modern syntax-aware textual editors support generation of code snippets via textual templates, triggered in a context-sensitive fashion by means of a content completion user interface. For example, Figure 1 demonstrates how a simple function is created using content completion.

In order to support the various applications of syntax definition, the different aspects of parsing, unparsing, formatting, and completion templates are often specified separately. For instance, Figure 2 shows an example of concrete syntax specification in Spoofax. Figure 2a defines a grammar, specified in SDF [8, 22], which is used to generate a parse table. Figure 2b defines completion templates that include layout and placeholder text. Figure 2c defines pretty printer specification that may lag behind modified code. Without going into the details of the three examples, it is immediately obvious that there is redundancy in these specifications.

Redundancy in syntax definitions is a common issue in tools to create syntax-aware textual editors, as we discuss in Section 7. It poses a maintenance problem as the language evolves: completion templates and pretty printer specification may lag behind modifications to the grammar of the language, hampering their completeness. One solution for the redundancy is to generate a default pretty printer from the syntax definition as applied for Spoofax in [10], and possibly generate default completion templates as well. Unfortunately, such a generative approach means that manually customized templates and formatting rules have to be combined with generated rules, which means it does not address the maintenance problem of these separate specifications.

---

\(^1\)http://strategoxt.org/Spoofax/TemplateLanguage
3. TEMPLATE-BASED SYNTAX DEFINITION

In this section we introduce a new syntax definition language based on templates as those found in template engines such as StringTemplate [14]. Additionally, we base the design of auxiliary features of the language, such as priority specification and lexical syntax, on that of SDF [8, 22]. The aim of the language is to eliminate the redundancy between different syntactic specifications, by combining concrete syntax, abstract syntax, formatting, whitespace, and placeholder names. We argue that through the use of templates, the language is rich in information yet elegant and simple.

Basic template-based syntax definitions consist of template productions that correspond to production rules in a grammar. They have the following form:

```source
s.label = <
  template
>
```

where `s` is the name of the symbol being defined, `label` is its constructor label used for the abstract representation, and `template` is a template that may include concrete syntax, references to other symbols (placeholders), and layout. Both the template and its placeholders are enclosed by `<...>` brackets.

As a first example, the following template productions define basic arithmetic expressions:

```source
templates
Exp.Num = <<INT>>
Exp.Plus = <<Exp> + <Exp>>
Exp.Times = <<Exp> * <Exp>>
```

The first production defines a template for number literals, defining a template for the `Exp` symbol based on a reference to the `INT` symbol. The other productions specify templates for the `+` and `*` operators. The last two templates consist of five elements: an `<Exp>` placeholder, whitespace, the `+` or `*` sign, more whitespace, and another placeholder. Of these elements, the whitespace elements are not considered for parser generation. Instead they are used for formatting in a generated pretty printer and completion templates. Whitespace characters treated this way are spaces, tabs, and newlines.

Placeholders can use the common `+` and `*` operators for repetition, and `*` for optionals. For repeated symbols with a separator symbol `s`, `<symbol`; separator=>`s` can be used. The following template productions illustrate these features, adding function calls and definitions to the expression language.

```source
templates
FunctionDef.Function = <
  function <ID>(<ID>; separator=",">) = <Exp>
>
Exp.Call = <<ID>(<Exp>; separator=",">>
```

**Disambiguation**

Grammars can be extended with disambiguation rules and annotations to express language characteristics such as associativity and operator precedence. In our running example, multiplication has a higher precedence than addition. Whereas in SDF [13] priorities are specified declaratively by copying the relevant productions and ordering them, separated by `>`-symbols, we add the option of specifying priorities through `references` to the relevant productions, so as to eliminate redundancy. The difference is shown in Figure 3.

**Lexical Syntax**

A part of syntax definitions we have not discussed so far is lexical syntax. Lexical syntax elements such as `INT` and `ID` in our expression language, are, unlike the context-free productions we discussed so far, generally specified using a form of regular expressions. In the abstract representation they are usually represented as simple strings, making unparsing trivial. For consistency, lexical productions can be specified in symbol-first order, as shown in Figure 4a. The definition of the body of lexical productions is shared with SDF [8, 22]. Both lexical and context-free syntax can be disambiguated using restriction sections in SDF [13], a construct that we inherit in our syntax specification language (Figure 4b).

The syntax of the template language is summarized in Figure 5. We proceed with a description of the mapping from syntax templates to SDF, completion templates, and pretty printing rules.

4. GENERATING TEMPLATE-BASED EDITORS

**Generating SDF**

Syntax templates closely match context-free syntax in SDF. Specifically, to go from a syntax template to an SDF production, all layout is discarded: in SDF, there is implicit `LAYOUT` between all symbols in a context-free production. The remaining elements (literals and placeholders) are converted in-order to their respective SDF equivalents. Layout is trimmed from the separator option of list placeholders. An example of this transformation is shown in Figure 6.

**Generating Completion Templates**

Completion templates in Spoofox consist of the following components:

- A symbol that indicates the context in which the template is applicable.
- A string, which is displayed in the completion pop-up, and is used to filter the list of proposals.
- A list of elements of the completion template. Each element is either a string, possibly including line breaks and indentation, a placeholder, or the special (cursor) directive. This last directive indicates the location of the cursor after the user
line that consists of a single ε-placeholder. Further instances of ε-placeholders on a single line are ignored: we expect the user to retrigger completion when they desire to insert a template at these positions. In Figure 7, \(<\text{MetaAnno}; \text{separator}\{\text{\textbackslash n}, \text{hide}\}\rangle\) is removed, and \(<\text{Statement}; \text{separator}\{\text{\textbackslash n}\}\rangle\) is replaced by the \(<\text{cursor}\rangle\) directive. Remaining ε-placeholders are replaced by an empty placeholder in the completion template that can be expanded to a single occurrence of one of the referred templates. In Figure 7 this is demonstrated by the replacement of \(<\text{FArg}; \text{separator}\{*, \text{\textbackslash n}\}\rangle\) by \(<\text{cursor}\>\).

### Generating Pretty Printing Rules

A simple set of recursive, bottom-up pretty printing rules can be generated from syntax templates. We generate these rules in the program transformation language Stratego. It can be seen in Figure 8 that a pretty printing rule consists of a number of components. The name of the rule, \(\text{pp\_Statement}\), is composed from the name of the symbol. The rule matches the constructor \(\text{ifThen}\) with two arguments. The number of arguments is equal to the number of placeholders in the syntax template.

When the rule matches, child nodes are pretty printed by (recursively) invoking (other) pretty printing rules. Then, the text for all elements of the template is concatenated, while the text for child nodes is indented by the amount the respective placeholder is indented in the syntax template.

### 5. Runtime Support for Template-Based Editors

For template-based editing to be effective, only contextually relevant templates should be shown. Pure structure editors achieve this based on the currently selected placeholder. To achieve the same in textual editors, the editor must be aware of the syntactic category of the text at the cursor location at any time. The pro-
Determining the syntactic category at the cursor location in a language-agnostic fashion is not trivial. If possible, changes to generated parsers support this facility should be provided. In addition, syntax errors need to be taken into consideration; the editor must be able to determine what type of template should be inserted even when a program is edited and is in a syntactically incorrect state.

**Original Implementation**

The solution originally implemented in Spoofax first creates a modified program text that includes a marker at the cursor location. The marker matches the syntax for identifiers, and is unlikely to be present anywhere else in the program text. The modified program text is then parsed, after which the AST is searched for the marker. Spoofax infers the symbols that should have been allowed at the position of the cursor from the token stream, and metadata attached to AST nodes.

The interaction between the involved components is a problem with this implementation. When the modified program text is parsed, it has parse errors, error recovery gets involved. Unless completion is invoked at a position where an identifier is allowed, there will be parse errors. The error recovery algorithm in the SGLR parser used in Spoofax attempts to get the parser back on track by performing a minimum number of token insertions and/or removals. As such, it may remove the marker, or insert punctuation that pushes the marker into another language construct.

One solution is to make the parser aware of the cursor location, and report the allowable syntactic categories at that character offset during parsing. This solution is specific to SGLR, and needs to be re-implemented in every other parser. We look for a more generic and less complex solution.

**Our Solution**

We apply a grammar generation technique that adds a production \( \text{CONTENTCOMPLETE} \rightarrow X (\text{COMP|UTION}-X) \) for every symbol \( X \), where \( \text{CONTENTCOMPLETE} \) is a symbol that recognizes the inserted marker text, including surrounding identifier characters. When the program text with marker is parsed, the marker can be parsed as every symbol allowed at its location. Because we encode the name of the symbol in the AST constructor, the editor runtime knows all symbols allowed at the position of the marker. The set of symbols is then used to select appropriate completion templates to display to the user.
The grammar of the Mobl language as specified by our template language is slightly more permissive than the original Mobl grammar, due to keywords that contain special characters, such as `@javascript` and `@doc`, which get tokenized by the SDF generator to `"javascript"` and `"doc"`, so that layout is allowed between these tokens. We will have to revisit this design decision, and consider, for example, removing the tokenization, and introducing a zero-length space character to insert `LAYOUT?`, to ensure it is possible to define such keywords in the template language.

Initially, syntactic completion in our evaluation was suboptimal due to the annotations present in Mobl at the start of many language constructs. Because the placeholder for these annotations is on a separate line, completion templates that produce a blank line before the language construct were generated. This behavior is likely not expected by the user, because these annotations are rarely used in Mobl. We corrected this by introducing the `hide` option to suppress the placeholders for annotations from the completion templates for many language constructs. The templates for annotations can be invoked separately, where desired.

**Evaluation of the Runtime Support**

We evaluated the runtime support for template-based editors on a sample program in the Mobl language by triggering completion on relevant positions in the sample program. Our approach was able to provide the editor runtime with an accurate and complete set of symbols on each sample position. The fact that our approach minimizes the interaction with error recovery likely explains these promising results.

7. RELATED WORK

**Unified Syntax Specifications**

There are a number of current syntax specification approaches that aim to unify the specification of parsing and unparsing.

SYN [3] aims to be one syntax definition language for the specification of ASTs, lexical analysis, parsing and pretty-printing. Its notation is similar to BNF, extended with a sublanguage for the specification of a lexical analyzer, and operators `h` (horizontal composition), `hV` (inconsistent line breaking), and `hov` (consistent line breaking) for the generation of a pretty printer. SYN has been implemented in Standard ML. Because the SYN compiler translates the syntax definition to input for the tools ML-Lex and ML-Yacc (ML implementations of the well known UNIX tools lex and yacc), a syntax definition in SYN faces the limitations of separate scanner and parser, and LALR(1) parsing.

Extended SDF [15] is an extension of SDF that embeds other specification languages. An important application of the work is the embedding of PP pretty printing rules (such as those in Figure 2c) in SDF attributes. Although this improves the locality of the different syntax definitions, it does not solve redundancy, as elements of the syntax are present both in the SDF, and in the attached pretty printing rule.

More recently, Rendel and Ostermann [16] propose partial isomorphisms for invertible computation, and use these to implement a combined parser/prett printer library in Haskell. Productions are specified using invertible combinators used for parsing, unparsing, and abstract syntax (de)construction.

These approaches differ from our approach in their use of explicit operators that specify layout and formatting. By using templates, we provide a concise syntax that forgoes the use of operators and uses plain whitespace instead, while still being sufficiently expressive for our case study with two complete DSLs. In addition, they also do not consider completion templates.

**Template-Based Editing**

Many early language workbenches used a template-based editing paradigm in structure editors. Examples include Centaur [2] and the Synthesizer Generator [17]. While these systems faced the same problem of having to specify both abstract and concrete syntax, they did not have the problem of specifying both a parser and an unparsers.

Hybrid textual/structure editors make it possible to switch to a text editing mode for a part of a program. Systems used to specify these editors do have the added dimension of parsing and unparsing, where they need a specification of formatted concrete syntax and a specification that specifies how to parse concrete syntax independent of the layout. While there have been different ways to address the issue, there has not been a solution that unifies the specification of all syntactic aspects. Examples of hybrid systems include the Programming System Generator (PSG) [1], PREGMATIC [21], and the ASF+SDF Meta-Environment [12].

PREGMATIC [21] specifies syntax using as part of attribute grammars. The grammar formalism does not include a formatting specification: instead, unformatted templates are generated from the grammar. The user can then edit the layout in those templates to format them as desired.

The Meta-Environment [12] uses SDF for syntax definition, and originally used the Generic Syntax-directed Editor (GSE) [4] as a hybrid editor. Contrary to many earlier structured editors it does not use pretty printing to convert abstract syntax into concrete syntax. Instead it maintains a two-way mapping between the text the user entered and the AST, so that a pretty printer is not needed during editing, and the user has full control over the layout of the program.

Template-based textual editors have text editing as their principal mode of operation, but can provide textual templates for editing. Examples of tools to create these editors include MontiCore [7], Xtext [5], and our own Spoofax [10]. Each of these systems has so far used a separate specification of syntax for parsing, pretty printing, and completion templates. As part of our work we implemented a template language for Spoofax, showing these aspects can be combined.

8. CONCLUSION

Syntax discoverability has been a crucial advantage of structure editors for new users of a language. Despite their excellent support for syntax discoverability, structure editors have not been widely adopted. Based on immediate parsing and analyses, modern textual code editors are also increasingly syntax-aware: structure and text-
tual editors are converging into a new editing paradigm that combines text and templates. In these syntax-aware textual editors, syntax discovery is provided in the form of syntactic completion, which depends critically on two features of a language workbench: first, it must be feasible to define and maintain sensible completion proposals for editors, and second, the syntactic completion proposals presented to the user must be relevant and complete. Current text-based languages require redundant specification of the different aspects that make up a template-based editor. Evolution of the language means that maintenance is required of all aspects in order to maintain completeness and consistency.

This paper addresses the issue of effective specification and implementation of syntax-aware textual editors. First, through unification of the specification of syntax, thus addressing the higher risk of incomplete/incorrect syntax due to redundant specifications. To accomplish this, we presented the design and implementation of a specification language that incorporates enough information so that syntax definition, pretty printer and completion templates can be generated. Doing so ensures completion templates are complete and up-to-date, thus improving the discoverability of syntax in the editor. Second, we describe techniques to accurately determine the syntactic categories at the cursor location, in order to present a relevant and complete set of completion proposals. We showed that by applying grammar generation techniques, it is possible to accomplish this goal without parser-specific modifications.

9. REFERENCES


