

Identifying Reusable Language Modules from Existing Domain-Specific Languages

David Méndez-Acuña, José A. Galindo, Benoit Combemale,
Arnaud Blouin, and Benoit Baudry

University of Rennes 1, INRIA/IRISA. France

{david.mendez-acuna,jagalindo,benoit.combemale,arnaud.blouin,benoit.baudry}@inria.fr

Abstract. The use of domain-specific languages (DSLs) has become a successful technique in the implementation of complex systems. However, the construction of this type of languages is time-consuming and requires highly-specialized knowledge and skills. In such context, a common practice for enabling reuse is the definition of languages modules that can be later put together in order to build up new DSLs. Nevertheless, the identification and reuse of those modules in a set of existing languages is currently relying on manual and costly tasks thus, hindering the reuse exploitation when developing DSLs. In this paper, we propose a solution to i) detect and identify possible reuse in a set of DSLs; and ii) break down those languages into modules that can be later exploited to minimize costs when constructing new languages and when maintaining the existing ones. Finally, we validate our approach by using a realistic DSLs coming out from industrial projects and obtained from public GitHub repositories.

1 Introduction

The use of domain-specific languages (DSLs) has become a successful technique to achieve separation of concerns in the development of complex systems [7]. A DSL is a software language in which expressiveness is scoped into a well-defined domain that offers a set of abstractions (a.k.a., *language constructs*) needed to describe certain system aspects. For example, we find DSLs for prototyping graphical user interfaces [24], specifying security policies [18] and or performing data analysis [10].

Naturally, the adoption of such language-oriented vision relies on the availability of the DSLs needed for expressing all the aspects of the system under construction [4]. This implies the development of many DSLs which is a challenging task due the specialized knowledge it demands. As a matter of fact, the ultimate value of DSLs has been severely limited by the cost of the associated tooling (i.e., editors, parsers, etc...) [13].

The research community in software languages engineering has proposed mechanisms to increase *reuse* during the construction of DSLs to reduce costs

when using DSLs. The idea is to leverage previous engineering efforts and minimize implementation from scratch [27]. These reuse mechanisms are based on the premise that “software languages are software too” [11] thus, it is possible to use software engineering techniques to facilitate their construction [14]. For instance, there are approaches that take ideas from Component-Based Software Engineering (CBSE) [5] and Software Product Lines Engineering (SPLE) [29] during the construction of new DSLs.

In practice, however, reuse is rarely achieved as a result of monolithic processes where language designers define language modules while trying to predict that they will be useful in future DSLs. Rather, the exploitation of reuse is often an iterative process where reuse opportunities are discovered during the construction of individual DSLs in the form of replicated functionalities that could be extracted as reusable language modules. For example, many DSLs offer expression languages with simple imperative instructions (e.g., `if`, `for`), variables management, and mathematical operators. Xbase [1] is a successful experiment that demonstrate that, using compatible tooling, such replicated functionality can be encapsulated and used in different DSLs.

The major complexity of this reuse process is that both, the identification of reuse opportunities and the extraction of the corresponding languages modules are manually-performed activities. Due the large number of language constructs within a DSL, and the dependencies among them, this process is tedious and error prone. Language designers must compare DSLs in order to identify commonalities and, then, to perform a refactoring process to extract those commonalities on separated and interdependent language modules. This idea stands on previous research shoulders that propose similar mechanisms for software in general [3].

In this paper we propose a computer-aided mechanism to automatically identify reuse opportunities within a given set of DSLs, and to extract reusable language modules from those commonalities. To this end, we define comparison operators that we use to detect commonalities in a given set of DSLs during a static analysis process. In turn, the extraction of the reusable language module is based on some basic principles from set theory. It is worth highlight that our approach is tool-supported; we provide an Eclipse-based IDE that implements the ideas presented in this paper. Besides, our approach considers not only the syntax of the DSLs but also their semantics.

The validation of our approach is twofold. On one hand, we evaluate *correctness* by demonstrating that our ideas and tool support can be actually applied in a real case study. In particular, we a case study composed of three different languages for modeling state machines that share certain commonalities [8]. We show that those commonalities can be extracted in reusable language modules. On the other hand, we evaluate the *relevance* of our approach. To this end, we explore public GITHUB repositories and download more than 2000 DSLs where we apply our analysis. Then, we found that there is a 40% overlapping within those DSLs representing potential reuse.

revisar mi escritura :)

The reminder of this paper is organized as follows: Section 2 introduces a set of preliminary definitions/assumptions that we use all along the paper. Section 3 presents a motivation to the problem by introducing an scenario. Section 4 describes our approach that is evaluated in Section 5. Section 6 presents the related work and, finally, Section 7 concludes the paper.

2 Background: Domain-specific languages in a nutshell

In this section we present some basic concepts that shows the scope of our approach.

- **Specification**→ Like general purpose languages (GPLs), DSLs are defined in terms of syntax and semantics [12]. Hence, the specification of a DSL is a tuple $\langle syn, sem, M_{syn \leftarrow sem} \rangle$ [6]. The parameter *syn* (the ***syntax***) refers to the structure of the DSL and specifies each language construct in terms of its name and the relationships it has with other language constructs. In turn, the parameter *sem* (the ***semantics***) refers to the meaning of the language constructs. This meaning corresponds to the dynamic behavior that establishes the manner in which language constructs are manipulated at runtime. Finally, the parameter $M_{syn \leftarrow sem}$ refers to the mapping between the language constructs and the semantics.
- **Technological space**→ Currently, there are diverse technological spaces available for the implementation of syntax and semantics of DSLs [21]. Language designers can, for example, choose between using context-free grammars or metamodels as specification formalism for syntax. Similarly, there are at least three technological spaces for expressing semantics: operationally, denotationally, and axiomatically [22].

In this paper we are interested on executable DSLs (xDSLs) which syntax is specified by means of metamodels, and semantics is specified operationally as methods (a.k.a, *domain-specific actions* [6]). Each language construct is specified in a metaclass. The relationships between language constructs are specified as references between metaclasses. In turn, domain-specific actions are specified as java-like methods that are allocated in the metaclasses.

Example: A DSL for finite state machines. Figure 1 shows a example xDSL for finite states machines. It contains thee classes `StateMachine`, `State`, and `Transition`; The class `StateMachine` contains both states and transitions which is represented with containment references. The code snippets at the bottom of Figure 1 introduce some operational semantics to this metamodel[6]. In our example, the metaclass `StateMachine` is enriched with the operation `eval()` that contains a loop that sequentially invokes the operation defined for the class `State`. This operation is also defined by using one aspect. The metaclass `Transition` is enriched with the operation `fire()`.

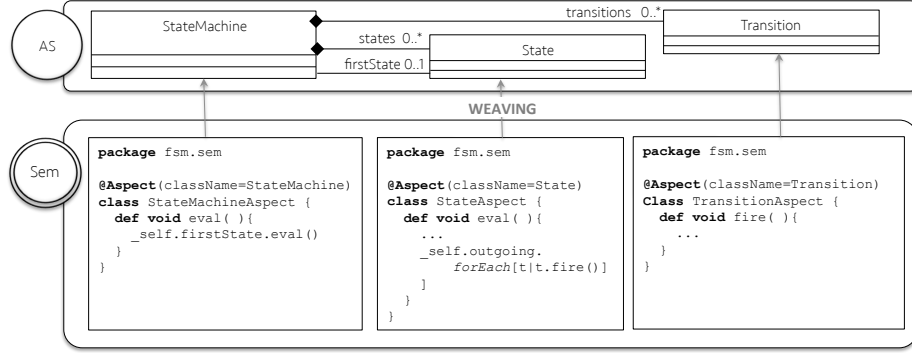


Fig. 1: A simple DSL for finite state machines

Este ejemplo debe ser algo que despues podamos usar para ejemplificar el funcionamiento del approach. Quizas, podamos mezclar esta sección con el motivating scenario y ahorrarnos una figura. Pero el ejemplo hay que usarlo para ejemplificar tu solución, no solo para saber que es un xdsl.

3 Motivation

Consider a team of language designers working on the construction of the DSLs for state machines presented in section 2. During that process, language designers implement the language constructs typically required for expressing finite state machines: states, transitions, events, and so on. In addition, the DSL is intended to provide a constraints language that allows final users to express guards on the transitions. Moreover, the DSL also provides an expressions language that allows to specify actions on the states of the state machine. This expressions language offers classical capabilities such as arithmetic operations, variables management, and imperative procedures.

After being released their DSL for state machines, the language development team is required again. This time the objective is to build a DSL for manipulating the traditional Logo turtle which is often used in elementary schools for teaching the first foundations of programming [23]. Certainly, the new DSL is essentially different from the DSL for state machines. Instead of states and transitions, Logo offers some primitives (such as **Forward**, **Backward**, **Left**, and **Right**) to move a character (i.e., the turtle) within a bounded canvas. However, Logo also requires an expressions language in order to specify complex movements. For example, final users may write instructions such as: **forward** ($x + 2$) where x is a variable.

At this point, language designers face the question of how to reuse the expressions language they already defined for the state machines language. As illustrated in Figure 2, the typical solution to this type of situations is to replicate the code in a second DSL. Language designers usually copy/paste the segment of

the specification that they can reuse. As a result, we have many clones that are expensive to maintain. Naturally, this situation is repeated each time that there is a new DSL to build. This fact is illustrated in the Figure 2 by introducing a third DSL for expressing flowcharts. In this case, the new DSL requires both, expressions and constraints languages.

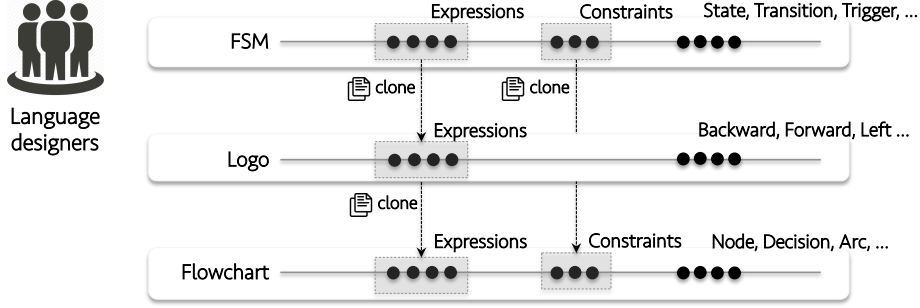


Fig. 2: Cloning in DSLs development process

Overlapping in DSLs and potential reuse: The aforementioned phenomenon was previously observed by Vöelter et al [28, p. 60-61]. In fact, that study demonstrates that although many of the existing DSLs are completely different and tackle independent domains; there are related DSLs with overlapping domains. That is, they share certain language constructs i.e., they have **commonalities** between them. Figure 3 illustrates this observation for the case of our illustrating scenario and by using two Venn diagrams to represent both syntax and semantic overlapping. Syntactic and semantic overlapping is represented as intersections between the corresponding sets. To this end, we designed an algorithm that is able to compute the all overlapping among the syntax of the DSLs in the input set.

If a set of DSLs have some overlapping and they are specified in the same technological space and using compatible language workbenches, then there is **potential reuse** since the specification of those shared constructs can be specified once and reused in the two DSLs [28, p. 60-61]. For the technological space discussed in this paper, syntactic commonalities appear where DSLs share some metaclasses and semantic commonalities appear where DSLs share some domain-specific actions.

It is worth to mention that the fact that two metaclasses are shared does not imply that all their domain specific actions are the same. We refer to that phenomenon as **semantical variability**. There are two constructs that share the syntax but that differ in their semantics. In such case, there is potential reuse at the level of the syntax since the metaclass can be defined once and reused in the DSLs but the semantics should be defined differently for each DSLs.

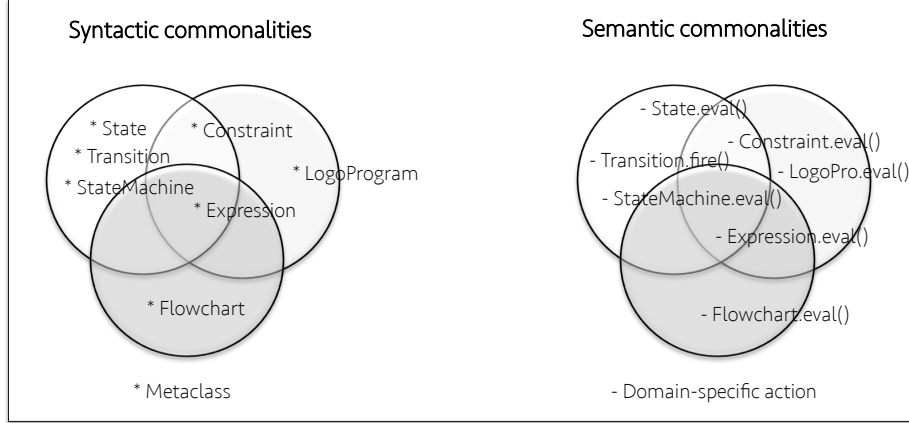


Fig. 3: Visualizing syntactic and semantic commonalities

4 Proposed approach

Given a set of existing DSLs our approach is intended to identify a catalog of reusable language modules from a given set of DSLs (that we refer to as the *input set*). To this end, our approach is threefold (see Figure 4). **First**, we identify commonalities existing in a given set of DSLs. Those commonalities are structured as the form of Venn diagrams containing syntax and semantic overlapping. **Second**, we break-down the Venn diagrams by separating all the intersections; for each different intersection, we create a language module by considering not only syntax but also semantics. **Third**, we encapsulate those language modules and we establish the dependencies among them by means of interfaces. The reminder of this section is dedicated to deeply explain our approach.

4.1 Identifying overlapping

Donde esta el algoritmo?? parece que cuentas de nuevo lo de arriba. Toca hablar de como comparas los lenguajes y generas el grafo para despues partirlo. Parece naive una cosa que no lo es. Se que lo haces por que eres buen profe, pero aqui hay que sacar pecho de lo super-mega-dificil que es.

Our algorithm for identifying **syntactic overlapping** can be described as by the function that receives a set of metamodels (one for each DSL of the input set) and returns a set of tuples containing all the overlapping (i.e., metaclasses that are *equal*) among these metamodels. Note that there can be overlapping among any of the combinations of the input set. Hence, in the result there is a tuple for each of the possible combinations of the input metamodels (i.e., the power set). Similarly, our algorithm for detecting **semantic overlapping** can be described as a function that receives a set of aspects (one for each DSL of

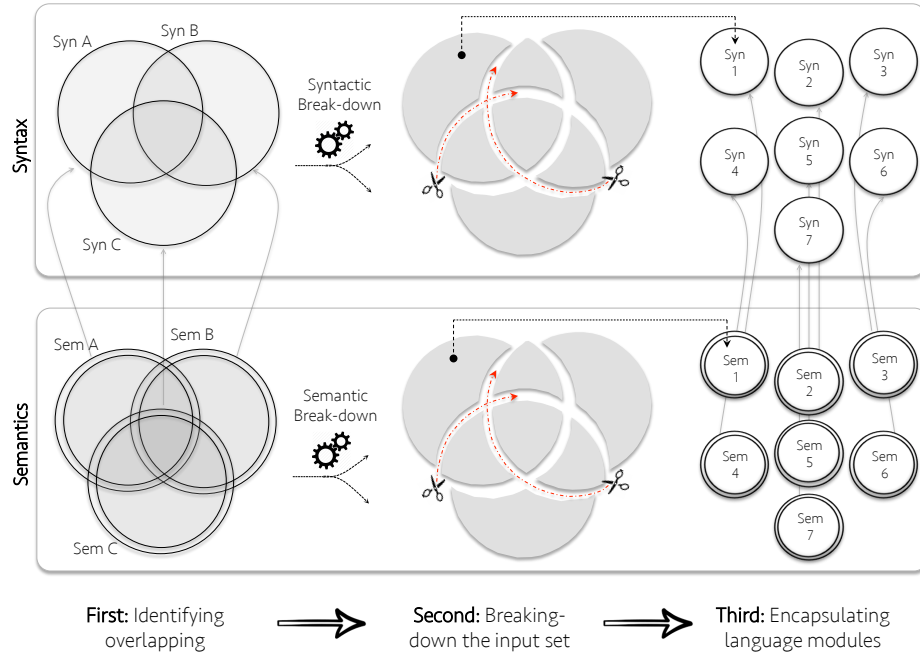


Fig. 4: Breaking down the input set by separating overlapping

the input set) and returns a set of tuples containing all the overlapping (i.e., domain-specific actions that are *equal*) among these aspects.

It is worth noting that a syntactic overlapping is a set of metaclasses that are equal in two or more DSLs. Similarly, a semantic overlapping is a set of domain-specific actions that are equal in two or more DSLs. At this point we need to clearly define the notion of equality between metaclasses and domain-specific actions. That is, we need to establish the criteria under we consider that two metaclasses/domain-specific actions are equal.

Comparison of metaclasses: The name of a metaclass usually corresponds to a word that evokes the domain concept the metaclass represents. Thus, intuitively one can think that a first approach to compare meta-classes is by comparing their names. Certainly, this approach results quite useful and it is quite probable that, we can find potential reuse.

me falta material. donde esta el fdx que representa el operador de igualdad. Explicame cuantos operadores hay, puedo tener mas? Tu solucion cuantos acepta?

Unfortunately, comparison of metaclasses by using only their names might have some problems. There are cases in which two meta-classes with the same name are not exactly the same since they do not represent the same domain concept or because there are domains that use similar vocabulary. In such cases,

an approach that certainly helps is to compare meta-classes not only by their names but also by their attributes and references.

Our comparison operator for metaclasses is quite restrictive. It identifies two metaclasses as a commonality if and only if their specifications are exactly the same. Hence, potential reuse is guaranteed in the sense that we are sure that there are not additional elements to consider. However, we are aware that there are other approaches to compute commonalities within metaclasses. So, at the implementation level our approach is flexible and additional comparison operators such as the surveyed in [17] can be easily incorporated.

toca usar el ejemplo con algunos operadores de ejemplo.

Comparison of domain-specific actions: Like methods in Java, domain-specific actions have a signature that specifies its contract (i.e., return type, visibility, parameters, name, and so on), and a body where the behavior is actually implemented. In that sense, the comparison of two domain-specific actions can be performed by checking if their signatures are equal. This approach is practical and also reflects potential reuse; one might think that the probability that two domain-specific actions with the same signatures are the same is elevated.

However, during the conduction of this research we realized that there are cases in which signatures comparison is not enough. Two domain-specific actions defined in different DSLs can perform different computations even if they have the same signatures. As a result, we only guarantee potential reuse where we compare also the bodies of the domain-specific actions. Note that such comparison can be arbitrary difficult. Indeed, if we try to compare the behavior of the actions we will have to deal with the semantic equivalence problem that, indeed, is known as be undecidable [19]. In this case, we a conservative approach is to compare only the structure (abstract syntax tree) body of the domain-specific action. To this end, we use the API for java code comparison proposed in [2].

Usa el ejemplo para ejemplificar

4.2 Breaking down the input set

After being identifying overlapping among the DSLs in the input set, we extract a set of reusable language modules. To this end, we adopt the idea presented by Vöelter et al [28, p. 60-61]; we break-down the overlapping by creating one language module for each different intersection as illustrated in Figure 4. The reasoning to this solution is quite simple: by definition, an intersection is a set of language constructs that are shared by two or more DSLs. If we extract those language constructs in separated language modules, the language module can be defined once and reused by all the DSLs that require it. So, we can consider that the extracted language module is reusable.

Usa el ejemplo para ejemplificar... no te preocupes por el espacio. Despues recortamos

4.3 Encapsulating language modules

Once we have identified the reusable language modules, we encapsulate them in such a way that each language module contains a metamodel and a set of domain-specific actions. Besides, the dependencies among the language modules are expressed.

The dependencies among language modules are expressed as follows: There is a *requiring module* that uses some constructs provided by a *providing module*. The requiring module has a dependency relationship towards the providing one that, in the small, is materialized by the fact that some of the classes of the requiring module have references (simple references, containment references, or inheritances) to some constructs of the providing one. In order to avoid direct references between modules, we introduce the notion of interfaces for dealing with modules' dependencies. The requiring language has a *required interface* whereas the providing one has the *provided interface*. A required interface contains the set of constructs required by the requiring module which are supposed to be replaced by actual construct provided by other module(s) (see Figure 5).

It is important to highlight that we use *model types* [?] to express both required and provided interfaces. A module can have some references to the constructs declared in its required interface. In turn, the relationship between a module and its provided interface is *implements* (deeply explained in [9]). A module implements the functionality exposed in its model type. If the required interface is a subtype of the provided interface, then the provided interface fulfills the requirements declared in a required interface.

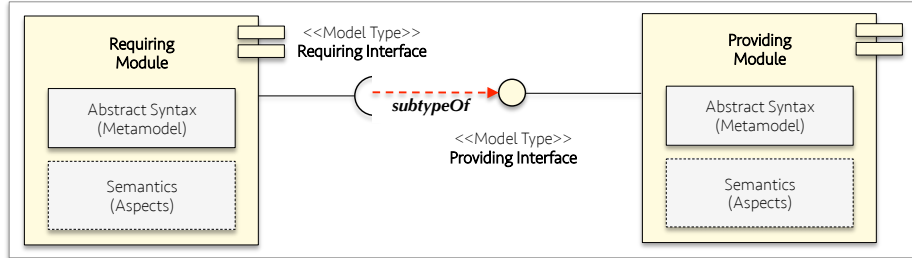


Fig. 5: Interfaces for language modules

Usa el ejemplo para ejemplificar... no te preocupes por el espacio. Despues recortamos... mira si esa figura puede ser mas pequena

5 Evaluation

The evaluation of our approach is twofold. First, we evaluate *correctness*. To this end, we use a test oracle consisting of a well-documented case study where we exactly know the existing overlapping among the involved DSLs. We execute our

approach on the case study and we check that the input matches the expected overlapping. Second, we evaluate *relevance*. More concretely, we use empirical data to demonstrate that the phenomenon of syntactic and semantic overlapping is actually appearing in realistic DSLs that we obtain from public GitHub repositories. Thus, we show that there is room in real projects to the applicability of our approach.

Implementation and hardware used → The approach presented within this paper is implemented in the puzzle toolsuite ¹. Puzzle have been developed as an Eclipse-based ² language workbench. Metamodels are specified in the Ecore language whereas domain-specific actions are specified as methods in Xtend programming language ³. The mapping between metaclasses and domain-specific actions is specified by using the notion of aspect introduced by the Kermeta ³ ⁴ that realize the ideas presented in [].

For the first experiment, Puzzle was executed in a Pc running MacOS X with XX Gb of ram and with a XXX Intel CPU. Then, for the second, a custom version of puzzle was extracted as an executable jar and executed within Grid5000 <https://www.grid5000.fr> to distribute the workload of analyzing all the possible overlapping within more than 2000 models.

5.1 Experiment 1: Evaluating *correctness* with the state machines case study

The case study we use to evaluate the correctness of our approach was introduced by Crane et al [8], and it is composed of three different DSLs for expressing state machines: UML state diagrams, Rhapsody, and Harel's state charts. This case study describes a realistic problematic that has an impact on the industry. As a matter of fact, it is one of the motivations of the VaryMDE⁵ project which is a collaboration between Thales Research & Technology, and INRIA.

The test oracle: Naturally, because the three DSLs are intended to express the same formalism, they have commonalities, namely, overlapping. For example, all of them provide basic concepts such as `StateMachine`, `State`, `Transition`, or `Trigger`. However, not all those DSLs are exactly equal. They have both syntactic and semantic differences which are well documented in the Crane's article [8].

Syntactic differences are materialized by the fact that not all the DSLs provide exactly the same constructs. Figure 6 shows a table with the language constructs provided by each DSL. There are differences in the support for transition's triggers and pseudostates. Whereas Rhapsody only supports simple triggers. Harel's state charts and UML provide support composed triggers. More concretely, in Harel's state charts triggers can be composed by using `AND`, `OR`,

¹ puzzle.github.io

² eclipse

³ <http://www.eclipse.org/xtend/>

⁴ <https://github.com/diverse-project/k3/wiki/Defining-aspects-in-Kermeta-3>

⁵ <http://varymde.gforge.inria.fr/>

cita tambien a melange si procede.

and NOT operators. In turn, in UML triggers can be composed by using the AND operator.

In addition, whereas there are pseudostates that are supported by all the DSLs (**Fork**, **Join**, **ShallowHistory**, and **Junction**); there are two pseudostates i.e., **DeepHistory** and **Choice** that are only supported by UML. The **Conditional** pseudostate is only provided by Harel’s state charts.

In summary, there are: **17** language constructs that are shared by all the DSLs; **1** construct that is exclusive of UML; **2** constructs that are exclusive of Harel’s state charts; **2** constructs shared by UML and Harel’s state charts; and, finally, **1** construct shared by Harel’s state charts and Rhapsody.

Language vs. Construct	StateMachine	Region	AbstractState	State	Transition	Trigger	NotTrigger	AndTrigger	OrTrigger	Pseudostate	InitialState	Fork	Join	DeepHistory	ShallowHistory	Junction	Conditional	Choice	FinalState	Constraint	Statement	Program	NamedElement
UML	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Rhapsody	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Harel	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Fig. 6: Oracle for evaluation of correctness

Semantic differences are materialized by the fact that not all the DSLs have the same behavior at execution time. For example, whereas in Harel’s state charts simultaneous events are attended in parallel, both UML and Rhapsody follow the run to completion principle so simultaneous events are attended sequentially.

As a direct consequence of those semantic differences, not all the domain-specific actions are the same. For instance, due to the semantic difference in the events treatment policy explained above, the methods `eval()` and `step()` in the **StateMachine** metaclass are different in each DSL.

Results: Let us now analyze the results we obtained from the execution of our approach on the test oracle. Figure 7 shows the results of the first part of the analysis. It presents the Venn diagram that shows syntactic and semantic overlapping. Note that, in the case of the syntactic overlapping, the cardinalities of the intersections in the Venn diagram match the test oracle thus, demonstrating that our approach to comparison of metaclasses is correct. In turn, the methods `eval()` and `step()` of the **StateMachine** metaclass are correctly identified as different in each DSL.

Figure 8 shows the results for the second and third steps of our approach: identifying and extracting reusable language modules. There is a language module that contains all the language constructs shared by the three DSLs. This language module can be considered as a basic DSL for state machines that supports the basic constructs. It can be used in the construction of new DSLs for

images/puzzle-overlapping.pdf

Fig. 7: Overlapping detected using the Puzzle toolsuite in the state machines case study.

state machines. In addition, there are other language modules encapsulate pseudostates and triggers separately. Note that in order to obtain the Harel's state charts language, we need to compose the modules 1, 2, and 5. In turn, to obtain UML we need to compose modules 1, 3, and 4. Finally, to obtain Rhapsody we need to compose modules 1 and 5.

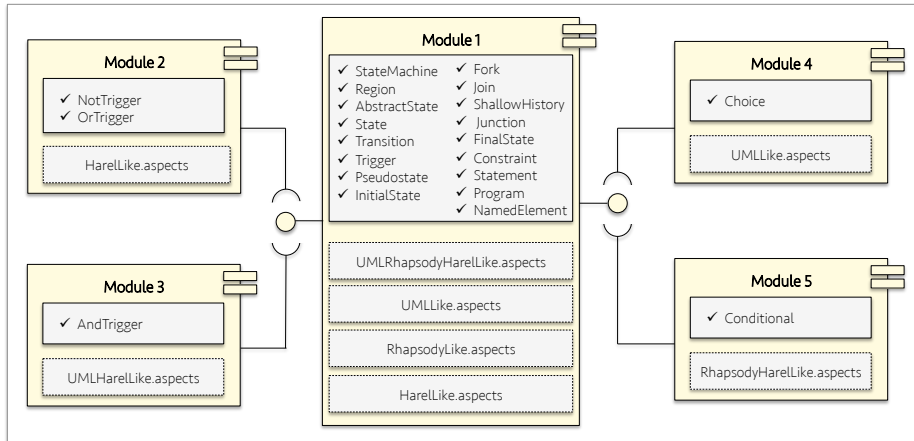


Fig. 8: Results for the state machines case study: extracting language modules

5.2 Experiment 2: Evaluating *relevance* by identifying potential reuse in the wild

In order to evaluate the relevance of our approach, we conducted an study on empirical data intended to answer two questions: 1) What is the probability that a DSL has some overlapping with another DSL?; and 2) How big is the average overlapping shared by the existing DSLs? The reminder of this section is dedicated to explain how we obtain the empirical data we use and how we use that data to answer the questions.

Collecting empirical data: To collect empirical data, we explored `GitHub` repositories in search of DSLs that are built on the same technological space that we used in our approach. Namely, metamodels written in `Ecore` with operational semantics defined as domain-specific actions in `Xtend`. The objective was to build a realistic data set composed of DSLs developed by diverse development teams.

As a result of this search, we found **2424** metamodels (after discarding metamodels with errors). Contrariwise, due to the fact that `Kermeta 3` and its implementation in `Xtend` is a quite recent idea, we found very few data for the semantics part.

We decided to conduct analysis only in the metamodels thus, the syntactic part of the languages. We consider that such analysis a good insight to know if there is potential reuse. In the following, we refer to our empirical data as S which is a set of metamodels.

What is the probability that a DSL has some overlapping with another DSL? To answer this first question, we compute the relative frequency of the event E : “the metamodel has overlapping with, at least, another metamodel” in the set S . Formally:

$$P(E) \approx \text{RelativeFrequency}(E) = \frac{|S_E|}{|S|} \quad (1)$$

$$\text{where } S_E = \{x \in S \mid (\exists y \in S \mid (x \neq y \wedge x \cap y \neq \emptyset))\}$$

Note that computing such probability corresponds to scan the set S in search of occurrences of the event E . After doing so, we obtained that the relative frequency is $974/2424 = 0.40$.

From a pragmatic point of view, this result means that there are **40%** of probabilities that, during the construction of a new DSL, a language designer is replicating at least one metaclass defined in a legacy DSL that is available in `GitHub`. This probability is quite elevated if we take into account that our comparison operator for metaclasses is quite restrictive.

Although this result is quite encouraging, we still wonder how important is the potential reuse. This motivate the following question:

How big is the average overlapping existing among DSLs? To compute the average of overlapping among the metamodels in the set S we built the

matrix illustrated in Figure 9. Basically, we performed a pair-wise comparison between the metamodels to obtain the overlapping between all the possible pairs. Then, we compute the average of overlapping that each metamodel has with the other metamodels of the set. To compute that average, we consider only the metamodels where the overlapping is at least 1 metaclass. Finally, we compute the average of these averages.

After executing this experiment, we obtain that the average of overlapping is 3.04. If we analyze the results of questions one and two at the same time, we can conclude that during the construction of the **40%** of the new DSLs language designers are replicating, in average, **3.04** language constructs.

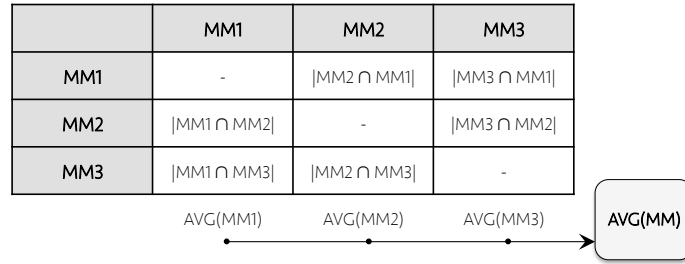


Fig. 9: Computing average commonalities

6 Related work

Software reuse has been largely studied during the last decades. The complexity behind the definition of reusable software is well-known and many approaches has been proposed to facilitate this task. The work of Caldiera and Basili [3] is a clear example of such approaches. It is intended to automatically extract reusable software modules from existing software systems. Then, a set of metrics for qualifying the quality of the produced modules. As mentioned in the introduction, this work has inspired the approach presented in that work paper.

The main contribution of the research presented this paper is that we apply those ideas in the construction of DSLs. To do so, we use a strategy based on Venn diagrams that, in turn, has been inspired in the phenomenon of domains overlapping identified by Vöelter et al [28, p. 60-61].

As a matter of fact, our approach is not the first one that tries to increase reuse in the construction of DSLs. The community of software languages engineering has been intensively working on this issue and nowadays we can find approaches for components-based languages development (such as [25,20,15]). In that context, our approach can be positioned as a reverse engineering technique for increasing reuse where reusable language modules are extracted from existing DSLs. It is worth noting that there are other approaches working on

reverse engineering for DSLs. For example, the research presented in [26] and [16], is intended to synthesize language product lines (i.e., software product lines where the products are DSLs) from a DSL specification.

7 Conclusions and future work

In this paper, we provide an approach for exploiting reuse during the construction of DSLs. We demonstrate that it is possible to partially automate the reuse process by identifying commonalities among DSLs and automatically extracting reusable language modules that can be later used in the construction of new DSLs. We evaluated our approach in a real industrial case study and we demonstrate that there is an important amount of potential reuse in DSLs in public repositories.

Acknowledgments

The research presented in this paper is supported by the European Union within the FP7 Marie Curie Initial Training Network “RELATE” under grant agreement number 264840; the VaryMDE project, a collaboration between Thales and INRIA; and the Cybersecurity project involving DGA (French department of defense) and INRIA. We also thank the support provided by the Grid5000 infrastructure during the execution of our experiments.

References

1. L. Bettini, D. Stoll, M. Völter, and S. Colameo. Approaches and tools for implementing type systems in xtext. In K. Czarnecki and G. Hedin, editors, *Software Language Engineering*, volume 7745 of *Lecture Notes in Computer Science*, pages 392–412. Springer Berlin Heidelberg, 2013.
2. B. Biegel and S. Diehl. Jccd: A flexible and extensible api for implementing custom code clone detectors. In *Proceedings of the IEEE/ACM International Conference on Automated Software Engineering*, ASE '10, pages 167–168, New York, NY, USA, 2010. ACM.
3. G. Caldiera and V. R. Basili. Identifying and qualifying reusable software components. *Computer*, 24(2):61–70, Feb. 1991.
4. T. Clark and B. Barn. Domain engineering for software tools. In I. Reinhartz-Berger, A. Sturm, T. Clark, S. Cohen, and J. Bettin, editors, *Domain Engineering*, pages 187–209. Springer Berlin Heidelberg, 2013.
5. T. Cleenewerck. Component-based dsl development. In F. Pfenning and Y. Smaragdakis, editors, *Generative Programming and Component Engineering*, volume 2830 of *Lecture Notes in Computer Science*, pages 245–264. Springer Berlin Heidelberg, 2003.
6. B. Combemale, C. Hardebolle, C. Jacquet, F. Boulanger, and B. Baudry. Bridging the chasm between executable metamodeling and models of computation. In K. Czarnecki and G. Hedin, editors, *Software Language Engineering*, volume 7745 of *Lecture Notes in Computer Science*, pages 184–203. Springer Berlin Heidelberg, 2013.

7. S. Cook. Separating concerns with domain specific languages. In D. Lightfoot and C. Szyperski, editors, *Modular Programming Languages*, volume 4228 of *Lecture Notes in Computer Science*, pages 1–3. Springer Berlin Heidelberg, 2006.
8. M. Crane and J. Dingel. Uml vs. classical vs. rhapsody statecharts: not all models are created equal. *Software & Systems Modeling*, 6(4):415–435, 2007.
9. T. Degueule, B. Combemale, A. Blouin, O. Barais, and J.-M. Jézéquel. Melange: A meta-language for modular and reusable development of dsls. In *8th International Conference on Software Language Engineering (SLE)*, Pittsburgh, United States, Oct. 2015.
10. J. Eberius, M. Thiele, and W. Lehner. A domain-specific language for do-it-yourself analytical mashups. In A. Harth and N. Koch, editors, *Current Trends in Web Engineering*, volume 7059 of *Lecture Notes in Computer Science*, pages 337–341. Springer Berlin Heidelberg, 2012.
11. J.-M. Favre, D. Gasevic, R. Lommel, and E. Pek. Empirical language analysis in software linguistics. In *Software Language Engineering*, volume 6563 of *LNCS*, pages 316–326. Springer, 2011.
12. D. Harel and B. Rumpe. Meaningful modeling: what's the semantics of “semantics”? *Computer*, 37(10):64–72, Oct 2004.
13. J.-M. Jézéquel, D. Méndez-Acuña, T. Degueule, B. Combemale, and O. Barais. When Systems Engineering Meets Software Language Engineering. In *CSD&M'14 - Complex Systems Design & Management*, Paris, France, Nov. 2014. Springer.
14. A. Kleppe. The field of software language engineering. In D. Ga?evi?, R. Lommel, and E. Van Wyk, editors, *Software Language Engineering*, volume 5452 of *Lecture Notes in Computer Science*, pages 1–7. Springer Berlin Heidelberg, 2009.
15. H. Krahn, B. Rumpe, and S. Völkel. Monticore: a framework for compositional development of domain specific languages. *International Journal on Software Tools for Technology Transfer*, 12(5):353–372, 2010.
16. T. Kühn, W. Cazzola, and D. M. Olivares. Choosy and picky: Configuration of language product lines. In *Proceedings of the 19th International Conference on Software Product Line*, SPLC '15, pages 71–80, New York, NY, USA, 2015. ACM.
17. L. Lafi, S. Hammoudi, and J. Feki. Metamodel matching techniques in mda: Challenge, issues and comparison. In L. Bellatreche and F. Mota Pinto, editors, *Model and Data Engineering*, volume 6918 of *Lecture Notes in Computer Science*, pages 278–286. Springer Berlin Heidelberg, 2011.
18. T. Lodderstedt, D. Basin, and J. Doser. Secureuml: A uml-based modeling language for model-driven security. In J.-M. Jézéquel, H. Hussmann, and S. Cook, editors, *UML 2002 - The Unified Modeling Language*, volume 2460 of *Lecture Notes in Computer Science*, pages 426–441. Springer Berlin Heidelberg, 2002.
19. D. Lucanu and V. Rusu. Program equivalence by circular reasoning. In E. Johnsen and L. Petre, editors, *Integrated Formal Methods*, volume 7940 of *Lecture Notes in Computer Science*, pages 362–377. Springer Berlin Heidelberg, 2013.
20. M. Mernik. An object-oriented approach to language compositions for software language engineering. *J. Syst. Softw.*, 86(9):2451–2464, Sept. 2013.
21. M. Mernik, J. Heering, and A. M. Sloane. When and how to develop domain-specific languages. *ACM Comput. Surv.*, 37(4):316–344, Dec. 2005.
22. P. D. Mosses. The varieties of programming language semantics and their uses. In D. Bjørner, M. Broy, and A. V. Zamulin, editors, *Perspectives of System Informatics*, volume 2244 of *Lecture Notes in Computer Science*, pages 165–190. Springer Berlin Heidelberg, 2001.
23. A. Olson, T. Kieren, and S. Ludwig. Linking logo, levels and language in mathematics. *Educational Studies in Mathematics*, 18(4):359–370, 1987.

24. S. Oney, B. Myers, and J. Brandt. Constraintjs: Programming interactive behaviors for the web by integrating constraints and states. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 229–238, New York, NY, USA, 2012. ACM.
25. E. Vacchi and W. Cazzola. Neverlang: A framework for feature-oriented language development. *Computer Languages, Systems & Structures*, 43:1 – 40, 2015.
26. E. Vacchi, W. Cazzola, B. Combemale, and M. Acher. Automating Variability Model Inference for Component-Based Language Implementations. In P. Heymans and J. Rubin, editors, *SPLC'14 - 18th International Software Product Line Conference*, Florence, Italie, Sept. 2014. ACM.
27. T. van der Storm, W. Cook, and A. Loh. Object grammars. In K. Czarnecki and G. Hedin, editors, *Software Language Engineering*, volume 7745 of *Lecture Notes in Computer Science*, pages 4–23. Springer Berlin Heidelberg, 2013.
28. M. Völter, S. Benz, C. Dietrich, B. Engelmann, M. Helander, L. C. L. Kats, E. Visser, and G. Wachsmuth. *DSL Engineering - Designing, Implementing and Using Domain-Specific Languages*. dslbook.org, 2013.
29. S. Zschaler, P. Sánchez, J. a. Santos, M. Alférez, A. Rashid, L. Fuentes, A. Moreira, J. a. Araújo, and U. Kulesza. Vml* a family of languages for variability management in software product lines. In M. van den Brand, D. Ga?evi?, and J. Gray, editors, *Software Language Engineering*, volume 5969 of *Lecture Notes in Computer Science*, pages 82–102. Springer Berlin Heidelberg, 2010.