

# Parser-Combinators for Context-Free Path Querying\*

Sophia Smolina  
Electrotechnical University  
St. Petersburg, Russia  
sofysmol@gmail.com

Ilya Kirillov  
Saint Petersburg State University  
St. Petersburg, Russia  
kirillov.iliya@gmail.com

Ekaterina Verbitskaia  
Saint Petersburg State University  
St. Petersburg, Russia  
kajigor@gmail.com

Semyon Grigorev  
Saint Petersburg State University  
St. Petersburg, Russia  
semen.grigorev@jetbrains.com

## ABSTRACT

Context-free path queries (CFPQs) allows one to formulate paths query in terms of context-free grammar and may be applied in such areas as biological data analysis and static code analysis. In order to simplify development of applications which use CFPQs it is necessary to provide transparent intergration of domain-specific languages for queries specification into general-purpose programming languages. Moreover it is necessary to provide static errors checking of queries. Such specific techniques as LINQ and ORM are used, but they have some problems with flexibility (for example, with query decomposition and subqueries reusing).

Parser combinators provide native and flexible way for language specification in terms of general-purpose language. While parser combinator are classical for linear input processing, we know only Trails library which apply this technique for path querying. Trails is limited because it has problems with left-recursive grammars and cycles handling. We show that it is possible to create general parser combinators for CFPQ which support arbitrary context-free grammar and arbitrary graph. We implement a library of such parser combinators and show that it is applicable for realistic problems.

## CCS CONCEPTS

• **Information systems** → **Graph-based database models; Query languages for non-relational engines;** • **Software and its engineering** → *Functional languages;* • **Theory of computation** → *Grammars and context-free languages;*

## KEYWORDS

Graph data bases, Language-constrained path problem, Context-Free path querying, Parser Combinators, Domain Specific Language, Generalized LL, GLL, Neo4J, Scala

### ACM Reference Format:

Sophia Smolina, Ekaterina Verbitskaia, Ilya Kirillov, and Semyon Grigorev. 2018. Parser-Combinators for Context-Free Path Querying. In *Proceedings*

\*This work is supported by grant from JetBrains Research.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

GRADES-NDA'18, June 2018, Houston, Texas USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 123-4567-24-567/08/06...\$15.00

[https://doi.org/10.475/123\\_4](https://doi.org/10.475/123_4)

of Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) 2018 (GRADES-NDA'18). ACM, New York, NY, USA, Article 4, 5 pages. [https://doi.org/10.475/123\\_4](https://doi.org/10.475/123_4)

## 1 INTRODUCTION

One useful type of graph queries is language-constrained path queries [3]. There are several languages for graph traversing/querying which support constraints formulated in terms of regular languages. For example SPARQL [18], Cypher<sup>1</sup>, and Gremlin [22]. In this work we are focused on context-free path queries (CFPQ) which use context-free languages for constraints specification and are used in bioinformatics [24], static code analysis [4, 17, 20, 28], and RDF processing [27]. There are a lot of theoretical research and problem-specific solutions on CFPQ [2, 8, 9, 16, 21, 24, 26], but there is a single known graph query language which supports CF constraints: cfsPARQL [27].

When one develops a data-centric application, one wants to use general purpose programming language and have a transparent and native access to data sources. String-embedded DSLs is one way to do it. It utilizes a driver to execute a query written as a string and to return a possibly untyped result. This approach has serious drawbacks. First of all, a DSL may require additional knowledge from a developer. Moreover, a string-embedded language itself is a source of possible errors and vulnerabilities static detection of which is very difficult [5]. In trying to solve these issues, such special techniques as Object Relationship Mapping (ORM) or Language Integrated Query (LINQ) [6, 15] were created. Unfortunately, they still experience difficulties with flexibility: for example with the query decomposition and the reusing of subqueries. In this paper, we propose a transparent and natural integration of CFPQs into a general-purpose language.

One natural way to specify a language is to specify its formal grammar which can be done by using special DSL based, for example, on EBNF-like notation [25]. The classical alternative way is a parser combinators technique which provide !!!Ekaterina, we need your help!!.

Unfortunately, classical combinators implement top-down parsing and cannot handle left recursive and ambiguous grammars [? ]. In [11] authors demonstrate a set of parser combinators which can

<sup>1</sup>Cypher language web page: <https://neo4j.com/developer/cypher-query-language/>. Access date: 16.01.2018

handle arbitrary context-free grammars by using ideas of Generalized LL [23] (GLL). Meerkat<sup>2</sup> parser combinators library is based on [11]. The result of parsing is represented in a compact form as Shared Packed Parse Forest [19] (SPPF). Paths extraction, queries debugging and result processing [10] require an appropriate representation of the query result. It is showed that SPPF is a suitable finite structural representation of a CFPQ query result, even if the set of paths is infinite [7].

An idea to use combinators for graph traversing has already been proposed in [13], but the solution presented provides only approximated handling of cycles in the input graph and does not support left-recursive grammars. Authors pointed out that the idea described is very similar to the classical parser combinators technique, but the supported language class or restrictions are not discussed.

In this paper we show how to compose these ideas and present the parser combinators for CFPQ which can handle arbitrary context-free grammars and provide structural representation of the result. We make the following contributions in this paper.

- (1) We show that it is possible to create parser combinators for context-free path querying which work on both arbitrary context-free grammars and arbitrary graphs and provide a finite structural representation of the query result.
- (2) We provide the implementation of the parser combinators library in Scala. This library provides an integration to Neo4J graph data base. Source code is available on GitHub: <https://github.com/YaccConstructor/Meerkat>.
- (3) We perform an evaluation on realistic data. Also we compare the performance of our library with another GLL-based CFPQ tool and with the Trails library. We conclude that our solution is expressive and performant enough to be applied to the real-world problems.

## 2 PARSE COMBITATORS FOR PATH QUERYING

Parser combinators provide a way to specify a language syntax in terms of functions and operations on them. A parser in this framework is usually a function which consumes a prefix of an input and returns either a parsing result or an error, if the input is erroneous. Parsers can be composed by using a set of parser combinators to form more complex parsers. A parser combinators library provides with a set of basic combinators (such as sequential application or choice), and there can also be user-defined combinators. Most parser combinators libraries, including the Meerkat library, can only process the linear input — strings or some kind of streams. We extend the Meerkat library to work on the graph input.

Meerkat library is a general parser combinators library; by using memoization, continuation passing style and the ideas of Johnson [12] it supports arbitrary context free specifications. This library is closely related to the Generalized LL algorithm and since GLL can be generalized for context free path querying [7], the adaptation of Meerkat is reasonable too [!!!  
ĐĖĐĤŃÄĐĤŃĐĐĤŃÄĐĤŃČĐĐŹĐŸŃÄĐĤĐšĐŹŃČŃŃ !!!]. It can be

done by using an odservation which for (string or graph) parsing we need only to provide function for getting symbols follower by specified position.

The combinators our library provides are presented in table 1. Basic parser combinators for mathching strings are implicitly generated whenever a string is used within a query. The same generation query can be written using the library as presented in Fig. 1.

Combinator	Description
$a \sim b$	sequential parsing: a then b
$a \mid b$	choice: a or b
$a.?$	optional parsing: a or nothing
$a.*$	repetition of zero or more a
$a.+$	repetition of at least one a

**Table 1: Meerkat combinators**

```
val S: Nonterminal = syn(
  "subclassof-1" ~ S.? ~ "subclassof" |
  "type-1" ~ S.? ~ "type")
```

**Figure 1: The same generation query in Meerkat**

The most exciting feature of our library is that queries can be used as first class values, which means greater generalization and comosition. Function sameGen presented in Fig 3 is a generalization of the same generation query which is independent from the environment such as the input graph structure or other parsers. It can be used for creation of different queries, including the one presented in Fig 1: it is the result of application of sameGen to the appropriate relations (which may be treated as “brackets”). Another application of the sameGen can be founded in Fig. 6.

```
val query1 = syn(sameGen(List(
  ("subclassof-1", "subclassof"),
  ("type-1", "type"))))
```

**Figure 2: Query ?? as an application of sameGen**

```
def sameGen(hrs) =
  bs.map { case (lbr, rbr) =>
    lbr ~ syn(sameGen(bs).?) ~ rbr }
  match {
    case x :: Nil => syn(x)
    case x :: y :: xs =>
      syn(xs.foldLeft(x | y)(_ | _))
  }
```

**Figure 3: Generic function for same generations query**

Running a query over an input graph retrieves the list of pairs  $(i, j)$  where each pair corresponds to the set of paths from the node

<sup>2</sup>Meerkat project repository: <https://github.com/meerkat-parser/Meerkat>. Access date: 16.01.2018

$i$  to the node  $j$ . Running the same generation query from Fig. 6 over the graph in Fig. 4 returns  $\{(1, 0), (1, 2)\}$  as a result. Internally in Meerkat this paths represented as SPPF [19]. Simplified version of SPPF for this query preseted on Fig. 5 where rounded rectangles represents nonterminals, others rectangles represent productions. Every rectangle contains nonterminal name or production rule, start and end nodes of thr input graph. Gray rectangles are start nonterminals.

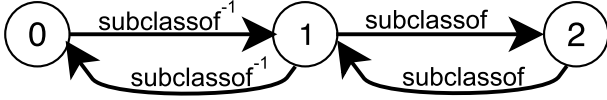


Figure 4: Example input graph.

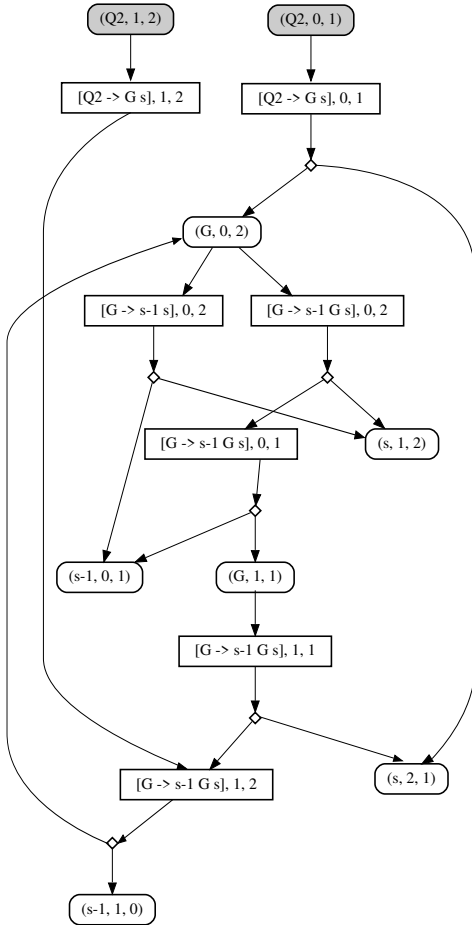


Figure 5: SPPF: result of application the same generation query 6 to the graph 4

### 3 EVALUATION

In this section we present evaluation of Meerkat grph querying library. We show its performance on a classical ontology graphs

for in memory graph and for Neo4j database, show application on may-alias static code analysis problem, and compare with Trails [13] library for graph traversals.

All tests are performed on a machine running Fedora 27 with quad-core Intel Core i7 2.5 GHz CPU with 8 GB of memory.

#### 3.1 Ontology querying

One of well-known graph querying problems is a queries for ontologies [1]. We use Meerkat to evaluate it on some popular ontologies presented as RDF files from paper [27]. We convert RDF files to a labeled directed graph like the following: for every RDF triple  $(subject, predicate, object)$  we create two edges  $(subject, predicate, object)$  and  $(object, predicate^{-1}, subject)$ . On those graphs we apply two queries from the paper [7] which grammars are in Fig. 2, and Fig. 6

```
val query2 = syn(
  sameGen(List(("subclassof-1", "subclassof"))) ~
    "subclassof")
```

Figure 6: Query 2 grammar

The queries applied in two following ways.

- Convert RDF files to a graph input for meerkat and then directly parse on query 1 and query 2
- Convert RDF files to a Neo4j database and then parse this database on given queries

Table 2 shows experimental results of those two aproaches over the testing RDF files where  $\#results$  is a number of pairs of nodes  $(v_1, v_2)$  such that exists S-path from  $v_1$  to  $v_2$ .

Meerkat and GLL [7] shows the same results (column  $\#results$ ) And if compare permomance to GLL on Query 1 Meerkat is a faster up to two times on some graphs (for example on measure-primitive), meanwhile on Query 2 GLL is faster.

In comparison of performance of in memory graph querying and database querying, the second one is slower in about 2 – 4 times.

#### 3.2 Static code analysis

Alias analysis is one of the fundamental static analysis problems [14]. Alias analysis checks may-alias relations between code expressions and can be formulated as a Context-Free language (CFL) reachability problem [20] which is closely the same as CFPQ problem. In that case program represeted as Program Expression Graph (PEG) [28]. Verticies in PEG are program expressions and edges are relations between them. In a case of analysing C source code there is two kind of edges **D**-edge and **A**-edge.

- Pointer dereference edge (**D**-edge). For each pointer deference  $*e$  there is a directed D-edge from  $e$  to  $*e$ .
- Pointer assignment edge (**A**-edge). For each assignment  $*e_1 = e_2$  there is a directed A-edge from  $e_2$  to  $*e_1$

Also, for the sake of simplicity, there are edges labeled by  $\bar{D}$  and  $\bar{A}$  which corresponds to reversed D-edge and A-edge, respectively.

The grammar for may-alias problem from [28] presented in Fig. 7. It consists of two nonterminals **M** and **V**. It allows us to make two kind of queries for each of nonterminals **M** and **V**.

Ontology	#triples	Query 1					Query 2				
		#results	In memory graph (ms)	DB query (ms)	Trails (ms)	GLL (ms)	#results	In memory graph (ms)	DB query (ms)	Trails (ms)	GLL (ms)
atom-primitive	425	15454	112	167	2849	232	122	49	52	453	19
biomedical-measure-primitive	459	15156	226	247	3715	482	2871	34	42	60	26
foaf	631	4118	16	25	432	29	10	1	2	1	1
funding	1086	17634	123	152	367	179	1158	18	23	76	13
generations	273	2164	6	21	9	12	0	0	0	0	0
people_pets	640	9472	63	84	75	80	37	2	3	2	1
pizza	1980	56195	544	650	7764	793	1262	44	47	905	50
skos	252	810	4	9	6	6	1	0	1	0	0
travel	277	2499	21	55	34	21	63	2	2	1	2
univ-bench	293	2540	15	43	31	24	81	2	2	2	1
wine	1839	66572	543	727	3156	606	133	5	7	4	5

Table 2: Comparison of Meerkat, Trails and GLL performance on ontologies

$$M \rightarrow \bar{D} V D$$

$$V \rightarrow (M? \bar{A})^* M? (A M?)^*$$

Figure 7: Context-Free grammar for the may-alias problem

Program	Code Size (KLOC)	Count of aliases		Time (ms)
		M aliases	V aliases	
wc-5.0	0.5	0	174	107
pr-5.0	1.7	13	1131	63
ls-5.0	2.8	52	5682	253
bzip2-1.0.6	5.8	9	813	71
gzip-1.8	31	120	4567	227

Table 3: Running may-alias queries on Meerkat on some C open-source projects

- **M** production shows that two l-value expression are memory aliases i.e. may stands for the same memory location.
- **V** shows that two expression are value aliases i.e. may evaluate to the same pointer value.

We made **M** and **V** queries on the code some open-source C projects. The results are presented on the Table 3

```
val M = syn("nd" ~ V ~ "d")
val V = syn((M.? ~ "na").* ~ M.? ~ ("a" ~ M.?).*)
```

Figure 8: Meerkat representation of may-alias problem grammar 7

### 3.3 Comparison with Trails

Trails [13] is a Scala graph combinator library. It provides traversers for describing paths in graphs in terms of parser combinators and allows to get results as a stream (maybe infinite) of all possible paths described by composition of basic traversals. Trails as well

as Meerkat support parsing in memory graphs, so we compare performance of Trails and Meerkat on the ontology queries which are described above. The result of comparison are presented in table 2. Trails gives the same results as Meerkat (column *results* in table 2) but slower than Meerkat up to 10 times on some graphs.

To summarise we show that parser combinators are expressive enough for a formulation of real queries. Performance of our implementation is better than performance of other existing combinators library and comparable with other GLL-based solution.

## 4 CONCLUSION

We propose a native way to integrate language for context-free path querying into general purpose programming language. Our solution can handle arbitrary context-free grammars and arbitrary graphs. Proposed approach is language-independent and may be implemented for closely all general-purpose programming languages. We implement it in Scala programming language and show that our implementation can be applied for real problems.

We can propose some possible directions of future work. First of all it is necessary to extend library with combinators for vertices information processing. The next technical improvement is creation of user-friendly interface for SPPF processing. For example, SPPF can be presented as a set of paths with additional information about its structure. It may be useful for SPPF utilization for debugging and query result processing.

Another direction is semantic actions (or attributed grammars) handling. It is useful for specification user-defined actions, such as filters, over subqueries result, which can make queries more expressive. It is impossible in general case, but some techniques such as lazy evaluations can help to provide a technically appropriate solution. An important theoretical question is for which class of semantic actions it is possible to provide precise general solution.

## REFERENCES

- [1] Serge Abiteboul, Richard Hull, and Victor Vianu. 1995. Foundations of Databases. (1995).
- [2] Pablo Barceló, Gaele Fontaine, and Anthony Widjaja Lin. 2013. Expressive Path Queries on Graphs with Data. In *International Conference on Logic for Programming Artificial Intelligence and Reasoning*. Springer, 71–85.

- [3] Chris Barrett, Riko Jacob, and Madhav Marathe. 2000. Formal-language-constrained path problems. *SIAM J. Comput.* 30, 3 (2000), 809–837.
- [4] Osbert Bastani, Saswat Anand, and Alex Aiken. 2015. Specification inference using context-free language reachability. In *ACM SIGPLAN Notices*, Vol. 50. ACM, 553–566.
- [5] Tevfik Bultan, Fang Yu, Muath Alkhalaf, and Abdalbaki Aydin. 2018. String Analysis for Software Verification and Security. (2018).
- [6] James Cheney, Sam Lindley, and Philip Wadler. 2013. A Practical Theory of Language-integrated Query. *SIGPLAN Not.* 48, 9 (Sept. 2013), 403–416. <https://doi.org/10.1145/2544174.2500586>
- [7] Semyon Grigorev and Anastasiya Ragozina. 2017. Context-free Path Querying with Structural Representation of Result. In *Proceedings of the 13th Central & Eastern European Software Engineering Conference in Russia (CEE-SECR '17)*. ACM, New York, NY, USA, Article 10, 7 pages. <https://doi.org/10.1145/3166094.3166104>
- [8] Jelle Hellings. 2014. Conjunctive context-free path queries. (2014).
- [9] Jelle Hellings. 2015. Path Results for Context-free Grammar Queries on Graphs. *CoRR abs/1502.02242* (2015). <http://arxiv.org/abs/1502.02242>
- [10] Piotr Hofman and Wim Martens. 2015. Separability by short subsequences and subwords. In *LIPICs-Leibniz International Proceedings in Informatics*, Vol. 31. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- [11] Anastasia Izmaylova, Ali Afroozeh, and Tijs van der Storm. 2016. Practical, General Parser Combinators. In *Proceedings of the 2016 ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation (PEPM '16)*. ACM, New York, NY, USA, 1–12. <https://doi.org/10.1145/2847538.2847539>
- [12] Mark Johnson. 1995. Memoization in Top-down Parsing. *Comput. Linguist.* 21, 3 (Sept. 1995), 405–417. <http://dl.acm.org/citation.cfm?id=216261.216269>
- [13] Daniel Kröni and Raphael Schweizer. 2013. Parsing Graphs: Applying Parser Combinators to Graph Traversals. In *Proceedings of the 4th Workshop on Scala (SCALA '13)*. ACM, New York, NY, USA, Article 7, 4 pages. <https://doi.org/10.1145/2489837.2489844>
- [14] Thomas J. Marlowe, William G. Landi, Barbara G. Ryder, Jong-Deok Choi, Michael G. Burke, and Paul Carini. 1993. Pointer-induced Aliasing: A Clarification. *SIGPLAN Not.* 28, 9 (Sept. 1993), 67–70. <https://doi.org/10.1145/165364.165387>
- [15] Erik Meijer, Brian Beckman, and Gavin Bierman. 2006. LINQ: Reconciling Object, Relations and XML in the .NET Framework. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)*. ACM, New York, NY, USA, 706–706. <https://doi.org/10.1145/1142473.1142552>
- [16] A. Mendelzon and P. Wood. 1995. Finding Regular Simple Paths in Graph Databases. *SIAM J. Computing* 24, 6 (1995), 1235–1258.
- [17] Polyvios Pratikakis, Jeffrey S Foster, and Michael Hicks. 2006. Existential label flow inference via CFL reachability. In *SAS*, Vol. 6. Springer, 88–106.
- [18] Eric Prud, Andy Seaborne, et al. 2006. SPARQL query language for RDF. (2006).
- [19] Joan Gerard Rekers. 1992. *Parser generation for interactive environments*. Ph.D. Dissertation. Universiteit van Amsterdam.
- [20] Thomas Reps. 1997. Program Analysis via Graph Reachability. In *Proceedings of the 1997 International Symposium on Logic Programming (ILPS '97)*. MIT Press, Cambridge, MA, USA, 5–19. <http://dl.acm.org/citation.cfm?id=271338.271343>
- [21] Juan L Reutter, Miguel Romero, and Moshe Y Vardi. 2015. Regular queries on graph databases. *Theory of Computing Systems* (2015), 1–53.
- [22] Marko A Rodriguez. 2015. The gremlin graph traversal machine and language (invited talk). In *Proceedings of the 15th Symposium on Database Programming Languages*. ACM, 1–10.
- [23] Elizabeth Scott and Adrian Johnstone. 2010. GLL parsing. *Electronic Notes in Theoretical Computer Science* 253, 7 (2010), 177–189.
- [24] Petteri Sevon and Lauri Eronen. 2008. Subgraph queries by context-free grammars. *Journal of Integrative Bioinformatics* 5, 2 (2008), 100.
- [25] Niklaus Wirth. 1996. Extended Backus-Naur Form (EBNF). *ISO/IEC 14977* (1996), 2996.
- [26] Mihalis Yannakakis. 1990. Graph-theoretic methods in database theory. In *Proceedings of the ninth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. ACM, 230–242.
- [27] Xiaowang Zhang, Zhiyong Feng, Xin Wang, Guozheng Rao, and Wenrui Wu. 2016. Context-free path queries on RDF graphs. In *International Semantic Web Conference*. Springer, 632–648.
- [28] Xin Zheng and Radu Rugina. 2008. Demand-driven Alias Analysis for C. In *Proceedings of the 35th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '08)*. ACM, New York, NY, USA, 197–208. <https://doi.org/10.1145/1328438.1328464>