Wrocław University
of Science and Technology

Faculty of Computer Science and Management
Field of study: Computer Science
Specialty: —

Engineering Thesis

FORMAL GRAMMAR PRODUCTION RULE PARSING TOOL

Karol Belina

keywords:
Parser combinators, context-free grammars, Extended Backus-Naur Form

short summary:

The thesis documents the process of designing and implementing a tool for parsing the production rules of context-free grammars in a textual form. It discusses the choice of Extended Backus-Naur Form notation over the alternatives and provides a mathematical model for parsing such a notation. The implemented parser can turn a high-level specification of a grammar into a parser itself, which in turn is capable of constructing a parse tree from arbitrary input provided to the program with the use of parser combinators.

Supervisor	dr inż. Zdzisław Spławski	....	....
	Title/degree/name and surname	grade	signature

The final evaluation of the thesis

Head of the examination commission	....	....	....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław 2020

Abstract

The thesis presents the design and implementation of a context-free grammar parsing tool with real-time explanations and error detection. It discusses the choice of Extended Backus-Naur Form notation over the alternatives and provides a mathematical model for parsing such a notation. For this purpose, the official specification of the EBNF from the ISO/IEC 14977 standard has been examined and transformed into an unambiguous form. A definition of a grammar is proposed to act as a result of the syntactic analysis phase formed with a technique called *parser combination*. A method of testing an arbitrary input against the language generated by the constructed grammar is described. The thesis shows the process of creating a simple command line REPL program to act as a basic tool for interfacing with the grammar parser and checker, but in order to efficiently use the library, a web-based application is designed on top of that to serve as a more visual, user-friendly and easily accessible tool. It describes the deployment of the application on a static site hosting service, as well as a cross-platform desktop application. The designed and implemented system gives the opportunity to extend it with other grammar specifications.

Streszczenie

Praca przedstawia proces projektowania i implementacji narzędzia służącego do analizy syntaktycznej gramatyk bezkontekstowych z naciskiem na obsługę błędów i wyjaśnień w czasie rzeczywistym. Omawia wybór rozszerzonej notacji Backusa-Naura i przestawia matematyczny model do analizy takiej notacji. W tym celu przeprowadzono analizę i przekształcenie w jednoznaczną formę oficjalnej jej specyfikacji zdefiniowanej w standardzie ISO/IEC 14977. Zaproponowana zostaje definicja gramatyki tej notacji, która jest tworzona w wyniku analizy syntaktycznej za pomocą techniki zwanej *kombinacją parserów*. Opisana zostaje metoda sprawdzania dowolnego ciągu znaków pod kątem języka generowanego przez analizowaną gramatykę. Praca przedstawia stworzenie prostego programu działającego z poziomu wiersza poleceń, który jest podstawowym narzędziem do analizy gramatyk, jednak by móc efektywnie korzystać ze stworzonej biblioteki, zaprojektowana zostaje aplikacja webowa, która służy za bardziej wizualne, przyjazne i łatwo dostępne dla użytkownika narzędzie. Praca opisuje wdrażanie aplikacji na usługę hostingową dla statycznych stron, a także jako wieloplatformowej aplikacji. Zaprojektowany i wdrożony system daje możliwość rozszerzenia go o inne specyfikacje gramatyk.

Contents

1. Problem analysis

1.1. Description and motivation

Programming language theory has become a well-recognized branch of computer science that deals with the study of programming languages and their characteristics. It is an active research field, with findings published in various journals, as well as general publications in computer science and engineering. But besides the formal nature of Programming language theory, many amateur programming language creators try their hand at the challenge of creating a programming language of their own as a personal project. It is certainly relevant for a person to write their own language for educational purposes, and to learn about programming language and compiler design. However, the language creator must first of all make some fundamental decisions about the paradigms to be used, as well as the syntax of the language.

The tools for aiding the design and implementation of the syntax of a language are generally called *compiler-compilers*. These programs create parsers, interpreters or compilers from some formal description of a programming language (usually a grammar). The most commonly used types of compiler-compilers are *parser generators*, which handle only the syntactic analysis of the language — they do not handle the semantic analysis, nor the code generation aspect. The parser generators most generally transform a grammar of the syntax of a given programming language into a source code of a parser for that language. The language of the source code for such a parser is dependent on the parser generator.

Most such tools, however, suffer from too much complexity and generally have a steep learning curve for people inexperienced with the topic. Limited availability makes them less fitted for prototyping a syntax of a language — they often require a complex setup for simple tasks, which is not welcoming for new users [**TODO** *and may lead to...?*]. The lack of visualization capabilities shipped with these tools makes them less desirable for teachers in the theory of formal languages, who often require such features for educative purposes in order to present the formulations of context-free grammars in a more visual format.

1.2. Goal of the thesis

The main goal of this thesis is to design and implement a specialized tool, that serves teachers, programmers and other kinds of enthusiasts of the theory of formal languages in the field of discrete mathematics and computer science, in order to formulate and visualize context-free grammars in the form of the Extended Backus-Naur Form. In order to [**TODO**], the tool must provide a graphical user interface. Additionally, to ensure the hightest degree of accessibility, the tool must be available in the form of an easily accessible web-based application that is accessed through a web page and can run in a browser without the need of installation on the user's device. The thesis itself will document the entire process of creating such a project.

[**TODO** *jak projekt pomoże w powyższych problemach?*]

In order to achieve the general goal, several sub-goals have been distinguished, all of which contribute to the main objective as a whole

- analysis of existing solutions and applications,
- presentation of the theoretical preliminaries of the project,
- definition of the outline of the project, including a description of the functional and non-functional requirements, the use case diagram, use case scenarios, the class diagram, and the user interface prototype,
- description of technologies used in the implementation,
- implementation of the project,
- description of the testing and deployment environments.

1.3. Scope of the project

The thesis will propose a definition of a grammar in the form of an abstract syntax tree of the Extended Backus-Naur Form. It will describe the process of implementing the business logic of the application in the Rust programming language compiled to WebAssembly. The compiled code is then ran inside the web-based application made with the Svelte framework, which incorporates the markup, CSS styles, and JavaScript scripts in the superset of the HyperText Markup Language (HTML).

The implementation phase will include the process of tokenization — the act of dividing the grammar in a textual form into a sequence of tokens — while taking into account proper interpretation of Unicode graphemes. The whitespace-agnostic tokens will be then combined together to form a previously-defined abstract syntax tree with a technique called *parser combination*. Several smaller helper parsers will be defined, all of which then will be combined into more sophisticated parsers capable of parsing entire terms, productions, and grammars. The implementation phase will also include the definition of an algorithm for handling left recursion in the resulting grammar, as well as a dependency graph reduction algorithm for determining the starting rule of a grammar. Up to this stage, any errors encountered in the textual form of a grammar are going to be reported to the user in a friendly format with exact locations of the errors in the input.

[TODO

- *service workers*
- *wizualizacje, edytor tekstowy i kolorowanie składni*
- *wyjaśnienia zwracane przez checker?*
- *wyrażenia regularne w specjalnych sekwencjach?*

]

The web application will be deployed on the GitHub Pages hosting service for static sites, as well as a standalone desktop application with the use of the Electron framework.

1.4. Glossary

AST Abstract syntax tree — **[TODO]**,

EBNF Extended Backus-Naur Form — **[TODO]**,

parser **[TODO]**,

REPL Read-Eval-Print loop — **[TODO]**.

[TODO]

2. Theoretical preliminaries

2.1. Formal grammars

2.1.1. Introduction to formal grammars

Formal grammar of a language defines the construction of strings of symbols from the language's *alphabet* according to the language's *syntax*. It is a set of so-called *production rules* for rewriting certain strings of symbols with other strings of symbols — it can therefore generate any string belonging to that language by repeatedly applying these rules to a given starting symbol [22]. Furthermore, a grammar can also be applied in reverse: it can be determined if a string of symbols belongs to a given language by breaking it down into its constituents and analyzing them in the process known as *parsing*.

For now, let's consider a simple example of a formal grammar. It consists of two sets of symbols: (1) set $N = \{ S, B \}$, whose symbols are *non-terminal* and must be rewritten into other, possibly non-terminal, symbols, and (2) set $\Sigma = \{ a, b, c \}$, whose symbols are *terminal* and cannot be rewritten further. Let S be the start symbol and set P be the set of the following production rules:

1. $S \rightarrow aBSc$
2. $S \rightarrow abc$
3. $Ba \rightarrow aB$
4. $Bb \rightarrow bb$

To generate a string in this language, one must apply these rules (starting with the start symbol) until a string consisting only of terminal symbols is produced. A production rule is applied to a string by replacing an occurrence of the production rule's left-hand side in the string by that production rule's right-hand side. The simplest example of generating such a string would be

$$S \underset{2}{\Rightarrow} \underline{abc}$$

where $P \underset{i}{\Rightarrow} Q$ means that string P generates the string Q according to the production rule i, and the generated part of the string is underlined.

By choosing a different sequence of production rules we can generate a different string in that language

$$\begin{aligned} S &\underset{1}{\Rightarrow} \underline{aBSc} \\ &\underset{2}{\Rightarrow} aB\underline{abc}c \\ &\underset{3}{\Rightarrow} a\underline{aB}bcc \\ &\underset{4}{\Rightarrow} aa\underline{bb}cc \end{aligned}$$

After examining further examples of strings generated by these production rules we may come into a conclusion that this grammar generates the language $\{ a^n b^n c^n \mid n \geq 1 \}$, where x^n is a string of n consecutive x's. It means that the language is the set of strings consisting of one or more a's, followed by the exact same number of b's, then followed by the exact same number of c's.

Such a system provides us with a notation for describing a given language formally. Such a language is a usually infinite set of finite-length sequences of terminal symbols from that language.

2.1.2. The Chomsky Hierarchy

In [5] Chomsky divides formal grammars into four classes and classifies them in the now called *Chomsky Hierarchy*. Each class is a subset of another, distinguished by the complexity.

Type-3 grammars generate the so-called *regular languages*. As described in [2], regular languages can be matched by *regular expressions* and decided by a *finite state automaton*. They are the most restricting kinds of grammars, with its production rules consisting of a single non-terminal on the left-hand side and a single terminal, possibly followed by a single non-terminal on the right-hand side. Because of their simplicity, regular languages are used for lexical analysis of programming languages [19].

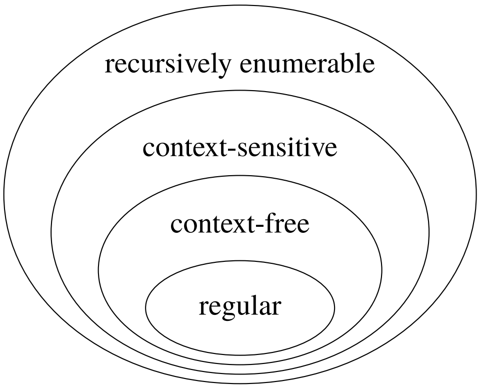


Figure 2.1: The Chomsky Hierarchy visualized

Type-2 grammars produce *context-free languages* and can be represented as a *push-down automaton* which is an automaton that can maintain its state with the use of a stack. [**TODO** *jak w stosie wygląda pamięć*]

[**TODO** *[17, 28]*]

2.1.3. Parsing Expression Grammars

[**TODO** `https://en.wikipedia.org/wiki/Parsing_expression_grammar`]
[**TODO** *[13]*]

2.2. Why EBNF?

[**TODO**]

2.3. Modifying the specification

[**TODO** *analiza i zmodyfikowanie oficjalnej specyfikacji EBNF*]
See appendix A.

2.4. Methods of syntactic analysis

[**TODO** *[1]*]

2.4.1. Bottom-up parsing

[**TODO**]

2.4.2. Top-down parsing and parser combination

[**TODO** *opisanie parser combinatorów (w Haskellu?) [34] [21] [12]*]

```
type Parser a = String -> Maybe (a, String)
```

3. Analysis of similar solutions

Coco/R

[**TODO** *[24]*]

ANTLR

[**TODO** *[25]*]

Bison

[**TODO** *[14]*]

PLY

[**TODO** *[3]*]

Regex101

[**TODO** *[9]*]

4. Design of the project

4.1. Requirements

4.1.1. Functional requirements

Table 4.1: The functional requirements of the project, their features, and priorities

Id	Requirement	Features	Priority
FR1	Specifying the grammar	The user can specify the grammar of a given language in the EBNF notation by providing it in a textual form in a designated editor window.	high
FR2	Error reporting	The editor provides feedback about any syntactic or semantic[1] errors encountered during the parsing by highlighting the exact location of the error in the provided grammar. The user can then hover the mouse pointer over the highlighted area to read the error message.	high
FR3	Specifying the input string	The user can specify the input string in a designated editor window to check if it belongs to the language generated by the previously-defined grammar.	high
FR4	Visualizing the parse tree	The application visualizes the parse tree resulting from parsing the specified input string with the parser generated by the grammar defined by the user.	high
FR5	Syntax highlighting	The editor highlights parts of the specified grammar with a different syntactic meaning in a different manner with the use of multi-colored fonts.	medium
FR6	Autocompletion of non-terminals	The editor predicts the identifier of a non-terminal a user is typing by providing a list of possible non-terminals, which then can be chosen by the user.	low
FR7	Production rule folding	The editor provides the ability to hide and reveal a production rule of the grammar inside the editor window.	low
FR8	Search and replace interface	The user can search for any occurrences of a phrase in the editor window and possibly replace them with a different phrase. The search and replace functionality should also support regular expressions.	low

[1] Such as production rule duplication or left recursion.

4.1.2. Non-functional requirements

Table 4.2: The non-functional requirements of the project and their priorities

Requirement	Priority
The web application should be available 24 hours a day, 7 days a week.	medium
Page loading time should be less than 1 second with internet download speed of 80 Mbps. Parsing and checking times should both be less than 50 milliseconds.	high
The application must work and display correctly in • Chrome version 86 or later, • Safari version 14 or later, • Edge version 86 or later, • Firefox version 82 or later, • Opera version 71 or later, as well as in the Electron framework version v10.1.5.	high
Usability **[TODO]**	medium
The source code of the product should be open source and freely available for possible modification and redistribution.	high
The project should include the documentation necessary for extension and maintenance of the system.	high
The system should provide high degree of integrability with future components which extend the functionalities of the system.	high

4.2. User stories

Table 4.3: The user stories

Id	User story
US1	As the user, I want to be able to paste the contents of my clipboard into the editor window in the application.
US2	As the user, I want to be able to type in the editor window with my keyboard.
US3	As the user, I want to be able to appreciate the multi-colored appearance of the text that represents the syntax that I provided.
US4	As the user, I want to be able to select a portion of the text in the editor window and copy it to the clipboard using a keyboard shortcut.
US5	As the user, I want to be able to hold the *Alt* key on my keyboard to create multiple cursors in the editor window.
US6	As the user, I want to have the ability to autocomplete the non-terminal I am typing that has already been declared elsewhere in the code.

US7	As the user, I want to be able to hide any existing production rules that might appear too long, to increase the degree of clarity and readability of the grammar I'm working on.
US8	As the user, I want to be able to show any previously hidden production rules of the grammar.
US9	As the user, I want to have the ability to press a certain key combination on my keyboard that would allow me to type a specific phrase in the popup window, which would then find all the occurrences of that phrase in the editor window.
US10	As the user, I want to be able to provide a regular expression for the *find* functionality that would allow me to find all occurrences of phrases that pattern match that specific regular expression.
US11	As the user, I want to be able to replace some of the occurrences of phrases found with the *find* functionality with another phrase provided in a popup window.
US12	As the user, I want to be able to specify the initial production rule in the process of checking the input string against the grammar I provided.
US13	As the user, I want to be able to see errors in the syntax of the provided grammar in the form of underlined text in the location of where the errors actually occur.
US14	As the user, I want to have the ability to hover the mouse pointer over the underlined text to read the error message at that location. Alternatively, I want to be able to hover over the error indicator, which appears next to the line number.
US15	As the user, I want to be able to see the parse tree of the recognized input string that I provided.
US16	As the user, I want to have the ability to collapse any nodes in the visualized parse tree that might appear too long.

4.3. Use case specification

4.3.1. Use cases

[**TODO** *redo this in visual paradigm*]

Figure 4.1: The use case diagram

Table 4.4: Descriptions of the use cases

Id	Name	Description
UC1	Specifying the grammar	Allows the user to specify the grammar of a given language in the EBNF notation by providing it in a textual form in a designated editor window.
UC2	Specifying the input string	Allows the user to specify the input string in a designated editor window to check if it belongs to the language generated by the previously-defined grammar.

UC3	Interacting with the visualization	Allows the user to observe the visualized parse tree of the provided input string and interact with it by expanding and collapsing the tree nodes.
UC4	Changing the initial rule	Allows the user to specify the initial production rule used in the process of checking the provided input string against the defined grammar.

4.3.2. Requirements traceability graph

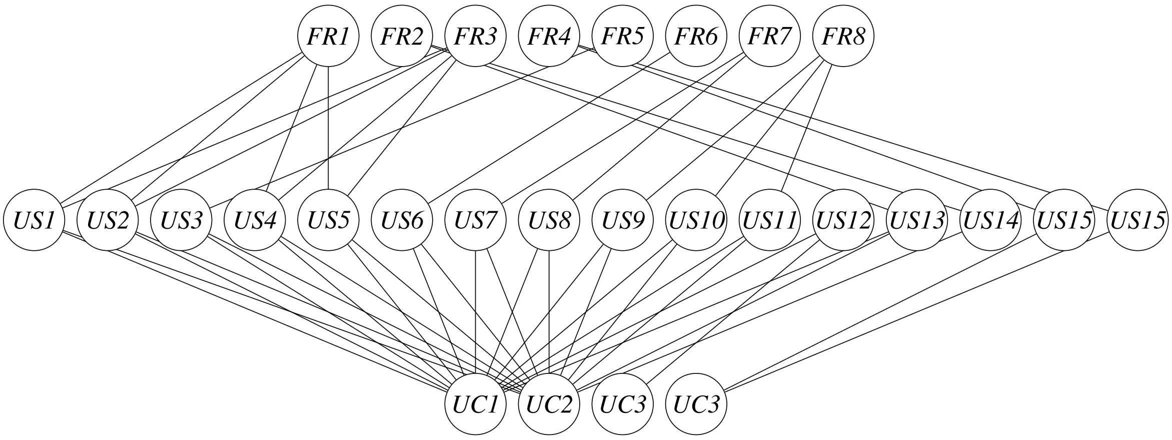


Figure 4.2: The requirements traceability graph

4.3.3. Use case scenarios

Table 4.5: Use case scenario of *UC1* Specifying the grammar

Identifier	*UC1*
Name	Specifying the grammar
Summary	Allows the user to specify the grammar of a given language in the EBNF notation by providing it in a textual form in a designated editor window.
Pre-conditions	None.
Post-conditions	The grammar has been correctly defined by the user with no syntactic errors.
Main scenario	1. The system shows a grammar editor window to the user. 2. The user provides a syntactically and semantically correct definition of a grammar. 3. The system shows an icon indicating no errors detected in the grammar. End of scenario.

Alternative scenario	2a.1. The user provides an invalid definition of a grammar. 2a.2. The system highlights the text in the grammar editor window at the error location. Return to step 2.

Table 4.6: Use case scenario of *UC2* Specifying the input string

Identifier	*UC2*
Name	Specifying the input string
Summary	Allows the user to specify the input string in a designated editor window to check if it belongs to the language generated by the previously-defined grammar.
Pre-conditions	None.
Post-conditions	The input string has been correctly entered by the user.
Main scenario	1. The system shows a input string editor window to the user. 2. The user provides a desired input string. 3. A valid grammar has been provided by the user in the grammar editor window. 4. The system shows the result of the checker in the result window. End of scenario.
Alternative scenario	3a.1. The user did not provide a valid grammar in the grammar editor window. 3a.2. The system does not show a result of the checker. End of scenario.

Table 4.7: Use case scenario of *UC3* Interacting with the visualization

Identifier	*UC3*
Name	Interacting with the visualization
Summary	Allows the user to observe the visualized parse tree of the provided input string and interact with it by expanding and collapsing the tree nodes.
Pre-conditions	The user has provided a valid definition of a grammar, as well as an input string, that belongs to the language generated by that grammar.
Post-conditions	None.
Main scenario	1. **[TODO]** 2. **[TODO]** 3. **[TODO]** End of scenario.

Table 4.8: Use case scenario of *UC2* Specifying the input string

Identifier	*UC4*
Name	Changing the initial rule
Summary	Allows the user to specify the initial production rule used in the process of checking the provided input string against the defined grammar.
Pre-conditions	The user has provided a valid definition of a grammar.
Post-conditions	The initial production rule has been successfully changed to the desired one.
Main scenario	1. The system shows a button the current initial production rule written on top. 2. The user clicks on the button. 3. The system shows a dropdown menu with a list of all production rules defined in the provided grammar. 4. The user clicks on an item of the list corresponding to the desired initial production rule. 5. The system changes the identifier of the initial production rule on the button. End of scenario.

4.3.4. Activity diagrams

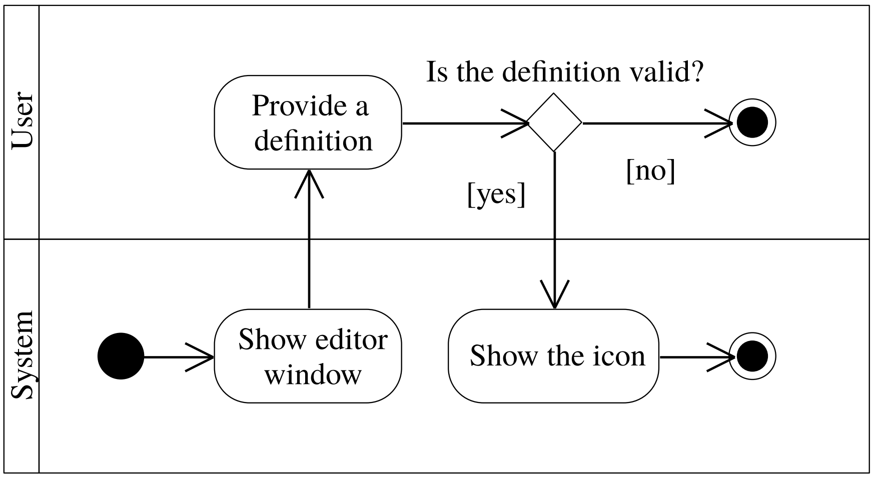


Figure 4.3: The activity diagram of *UC1* Specifying the grammar

4.3.5. Sequence diagrams

[**TODO** *redo this in visual paradigm*]

Figure 4.4: The sequence diagram of *UC1* Specifying the grammar

4.4. System architecture

4.4.1. Logical architecture

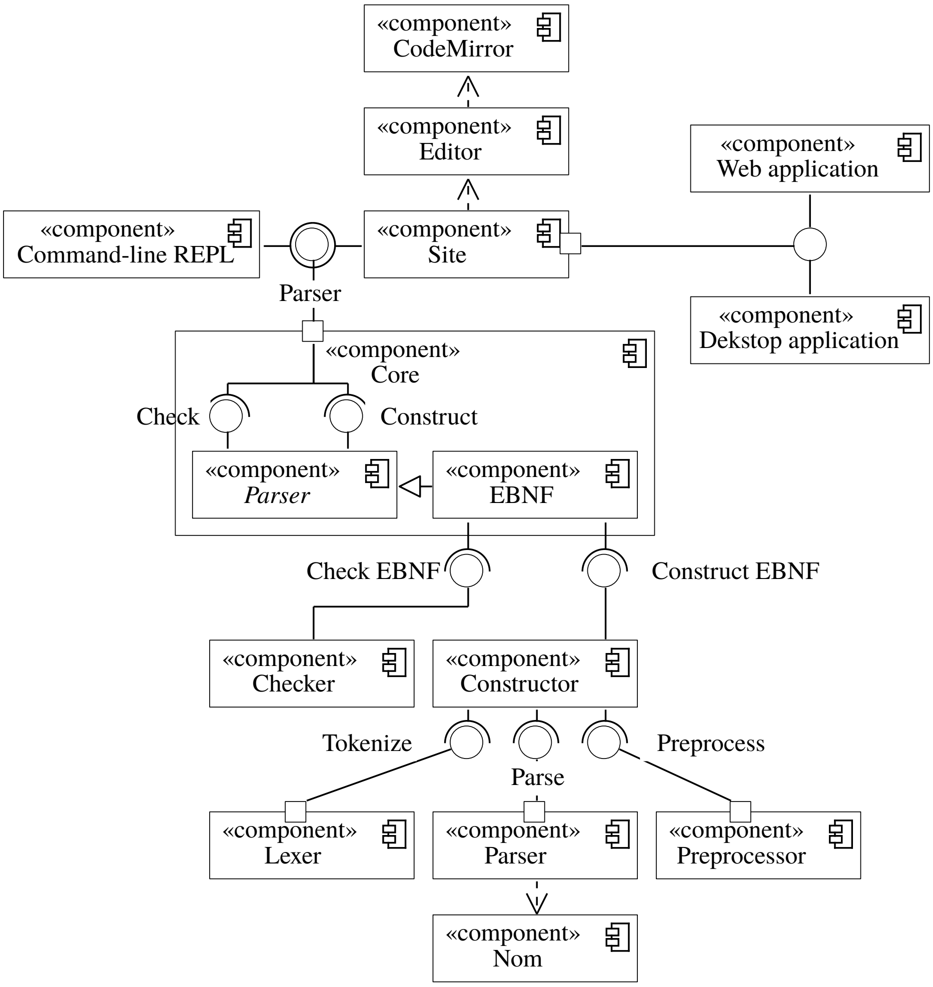


Figure 4.5: The logical architecture of the system represented with a UML component diagram

4.4.2. Physical architecture

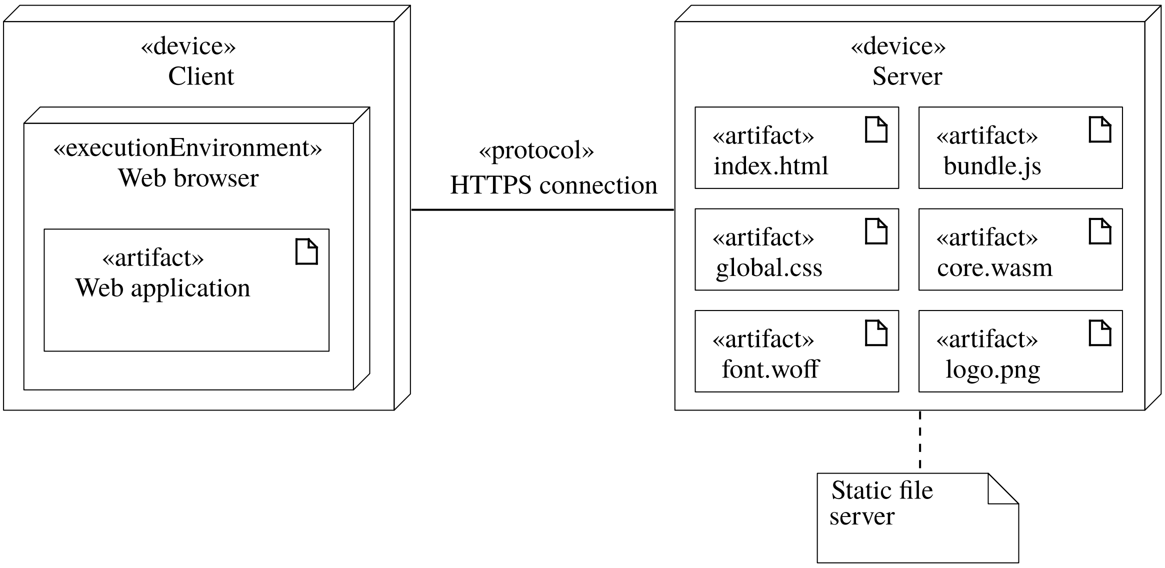


Figure 4.6: The physical architecture of the system represented with a UML deployment diagram

4.5. Interface prototype

[**TODO** *obrazki*]

5. Implementation of the project

5.1. Software environment

5.1.1. Technology infrastructure

[**TODO** *użyte technologie i zwizualizowany stack*]

[**TODO** *Git*] [**TODO** *Rust [20]*] [**TODO** *nom [7]*] [**TODO** *Svelte [31]*] [**TODO** *Rollup*] [**TODO** *WebAssembly*] [**TODO** *pakiety npm i cargo?*]

[**TODO** *opis technologii*]

5.1.2. Software

Visual Studio Code

Visual Studio Code [23] is a free, open-source text editor made by Microsoft for Windows, Linux and macOS. It is designed to write code and features syntax highlighting, code completion, snippets, code refactoring, and code debugging. The editor can be used with various programming languages, and supports extensions, which can be installed through a central repository called VS Code Marketplace available in the editor itself. The extensions may provide feature additions to the editor, as well as the support for various programming languages in the form of code linters, static code analysers, and debuggers. The editor is integrated with various version control systems, including Git and Subversion

According to the 2019 Developers Survey of Stack Overflow, Visual Studio Code ranked #1 among the top popular developer tools, with 50.7 % of the 87317 respondents using it. [30]

The extensions for the editor are created by the members of Visual Studio Code community. Two main extensions used by the author to develop the project were:

rust-analyzer [11] An implementation of the Language Server Protocol for the Rust programming language, which provides features such as code completion, messages for syntax and semantic errors, code actions, diagnostics, “go to definition” and other editor actions.

Svelte for VS Code [33] An implementation of the Language Server Protocol for the Svelte framework. The extension provides diagnostic messages for warnings and errors, support for Svelte pre-processors that provide source maps, as well as the support for Svelte-specific formatting (via prettier-plugin-svelte). Besides the Svelte language, the extension supports features such as hover info, messages for syntax and lint errors, and autocompletions for HTML, CSS/SCSS/LESS, as well as TypeScript and JavaScript.

The extensions has not proven to be crucial for the development of the project, but were an excellent addition to the workflow.

Besides the editor extensions, the terminal integrated with Visual Studio Code editor has been a valuable feature throughout the development process. The command line is a substantial factor in the development of modern applications, so a built-in terminal window allows the user to swiftly switch between the code editor and the command line.

The support for the Git version control system has also been advantageous when it comes to code editing. Every added, modified, or removed line of code is highlighted with an appropriate color in the code editor. This greatly improves the readability of the code, and allows the users to revert the code to its previous state right from the editor without any external tools.

Git

Git [15] is a free and open source distributed version control system. It has been a major part of the development process for the project, and has been used mainly as a tool for keeping track of the changes made to the source code and for integrating features in a smooth, non-disruptive manner.

Git supports branching and merging, which means that several project features may be implemented simultaneously and independently on separate *branches* and then *merged* into the main project. Every major code change has been implemented on a designated branch and was merged into the main branch only after a thorough testing process — this has made parallel development very easy, by isolating new development from finished work. This style of a workflow is known as GitFlow, made popular by Vincent Driessen [10], it has shown itself to be very effective for projects of any scale. Efficient switching between different versions of project files enables developers to work effectively on the project. Git includes specific tools for visualizing and navigating a non-linear development history. The author used [4] as a reference for using the tool.

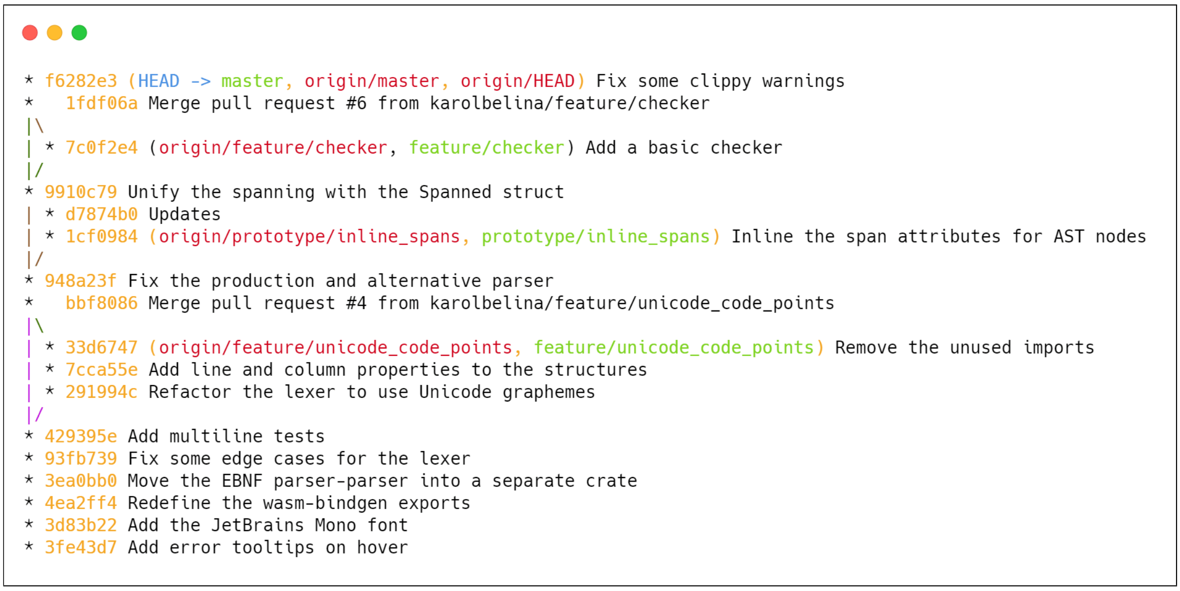
Git is now the most widely used source-code management tool, with 87.2 % of the 74298 respondents of the 2018 Developers Survey of Stack Overflow reporting that they use Git as their primary source-control system. [29].

The main client of Git used in the project was the command-line tool on the Ubuntu operating system running on Windows Subsystem for Linux. Figure 5.1 shows an example of GitFlow's *feature branches* and changes in the project repository in the Git version control system.

GitHub

GitHub [16] is a for-profit company owned by Microsoft that offers a cloud-based Git repository hosting service. As a company, GitHub makes money by selling hosted private code repositories, as well as other business-focused plans that make it easier for organizations to manage team members and security. The author used the free GitHub plan as the service for hosting the project's Git repository. Having the source code on an external server protected the project against data loss and allowed the developer to work on the project from any device at any convenient time.

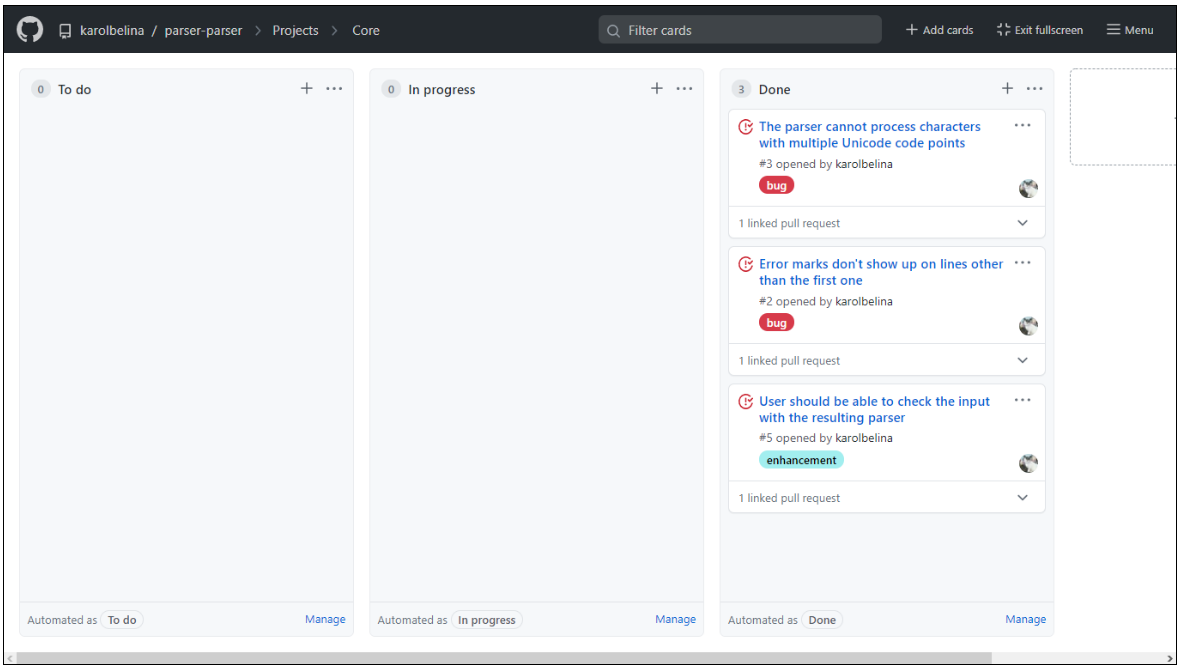
In addition to using GitHub as a hosting service, one can also exploit its project management features. Developers can create project boards related to the project's code repository, which are simple kanban board that can help organize and prioritize the work. With projects, the developers have the flexibility to manage boards for an entire project, or just for specific features. Figure 5.2 shows an example project board from Parser-parser.



```
* f6282e3 (HEAD -> master, origin/master, origin/HEAD) Fix some clippy warnings
*   1fdf06a Merge pull request #6 from karolbelina/feature/checker
|\
| * 7c0f2e4 (origin/feature/checker, feature/checker) Add a basic checker
|/
* 9910c79 Unify the spanning with the Spanned struct
| * d7874b0 Updates
| * 1cf0984 (origin/prototype/inline_spans, prototype/inline_spans) Inline the span attributes for AST nodes
|/
* 948a23f Fix the production and alternative parser
*   bbf8086 Merge pull request #4 from karolbelina/feature/unicode_code_points
|\
| * 33d6747 (origin/feature/unicode_code_points, feature/unicode_code_points) Remove the unused imports
| * 7cca55e Add line and column properties to the structures
| * 291994c Refactor the lexer to use Unicode graphemes
|/
* 429395e Add multiline tests
* 93fb739 Fix some edge cases for the lexer
* 3ea0bb0 Move the EBNF parser-parser into a separate crate
* 4ea2ff4 Redefine the wasm-bindgen exports
* 3d83b22 Add the JetBrains Mono font
* 3fe43d7 Add error tooltips on hover
```

Figure 5.1: Screenshot of the command-line interface of the Git version control system.

Project boards contain *issues* and *pull requests*, which can be moved from one kanban column to another, indicating that some work is currently "to do", work in progress, or complete. These work "cards" contain information about the author, assignees, the status, as well as simple textual notes. The *issues* are a way of reporting ideas, bugs, enhancements, or tasks natively on GitHub. After completing the work on an issue, a developer might create a *pull request* to allow other developers on the project to review and discuss the changes made to the code, and then deploy the changes by "pulling" the code to the central code repository.



0 To do

Automated as To do — Manage

0 In progress

Automated as In progress — Manage

3 Done

The parser cannot process characters with multiple Unicode code points
#3 opened by karolbelina
bug
1 linked pull request

Error marks don't show up on lines other than the first one
#2 opened by karolbelina
bug
1 linked pull request

User should be able to check the input with the resulting parser
#5 opened by karolbelina
enhancement
1 linked pull request

Automated as Done — Manage

Figure 5.2: Screenshot of one of the project's kanban boards on GitHub. [**TODO** *zaktualizować zrzut*]

GitHub supports Continuous integration and Continuous Delivery functionalities in form of *Actions* and *Pages*. GitHub Actions are a way to automate and execute any software development workflow after any change to the code in the repository. The user may set up many various actions for testing the changes on many development environments and operating systems at the same time, as well as building and deploying the code as a package or an arbitrary artifact. An action consists of jobs, which are defined by a list of steps required to execute them.

The GitHub Actions are used by the author to automate the testing and build process on every change made to the code repository. The built application is then deployed to a static site hosting service called GitHub Pages, which integrates itself seamlessly with Actions and GitHub repositories. GitHub Pages allows the user to host a website directly from a GitHub repository by combining static HTML, CSS, JavaScript, and other files straight from a repository into a website and publishing it on a `github.io` domain or a custom one.

Rust

Rust [27] is the main programming language used in Parser-parser — it powers the business logic part of the project. The language has been the most loved language for four years in a row in the Stack Overflow's survey [30]. The core idea of the language is memory safety — the language enforces certain rules checked at compile time, which guarantee that the program is safe from bugs like dereferencing null or dangling pointers, as well as making it difficult for the programmer to leak memory. Rust does this through a system of ownership and borrowing. The language, besides the safety, focuses on speed — its design lets the developer create programs that have the performance and control of a low-level language, but with the powerful abstractions of a high-level language.

Rust's design borrows heavily from the one of Haskell — both languages feature a rich type system, both are immutable-by-default, avoid mutation of shared references et cetera. Many developers tend to write Rust code in a functional style and adhere to the principles of functional programming, even though the language is multi-paradigm.

Without the need of a garbage collector, Rust projects are well-suited to be used as libraries by other programming languages. The language over the last few years has manifested itself in several distinct domains, including command line tools, networking, and embedded systems. Rust is supported on multiple operating systems and targets multiple platforms, has notable documentation, a user-friendly compiler with convenient error messages, and excellent tooling and ecosystem. For referencing the language, the author used [20], which covers many features and concepts of Rust.

WebAssembly

WebAssembly [36] (abbreviated *Wasm*) is a safe, portable, low-level code format designed for efficient execution and compact representation. Its main goal is to enable high performance applications on the Web, working alongside JavaScript, but not to be a replacement of it. It is designed to be portable, compact, and execute at or near native speeds. Although it has currently gathered attention in the JavaScript and Web communities in general, Wasm makes no assumptions about its host environment. WebAssembly is supported as a target for many programming languages, including C♯ via Blazor, C++ via EmScripten, and the main language used in Parser-parser — Rust. The author compiles Rust code to WebAssembly to be then used in a web environment for several reasons:

- Code size is incredibly important since the `.wasm` file must be downloaded over the network. Rust lacks a runtime, enabling small Wasm sizes because there is no extra code included, like a garbage collector,

- Rust and WebAssembly integrates with existing JavaScript tooling. It supports ECMAScript modules and the developer can continue using the tooling their already use, like npm and Webpack,
- JavaScript Web applications struggle to attain and retain reliable performance. The code is required to be ran frequently, so Wasm can solve this kind of problem with better memory and CPU efficiency at a lower level compared to the JavaScript interpreter.
- The Rust language itself, with a strong package manager, high performance, memory safety, and zero-cost abstractions.

Cargo

Cargo is the Rust's package manager. It downloads the Rust package's dependencies and compiles them, ensuring that the developer will always get a repeatable build. To accomplish this goal, Cargo introduces two metadata files with various bits of package information, fetches and builds the dependencies, invokes the Rust compiler with correct parameters to build the package, and introduces conventions to make working with Rust packages easier.

Rust provides first-class support for unit and integration testing, and Cargo allows the developer to execute all tests with a single command. Additionally, Cargo allows the developer to install extensions, which enhance the workflow and the development process. One of extensions useful for the author was Clippy — a collection of lints to catch common mistakes and improve the Rust code.

`crates.io` is the Rust community's central package registry that serves as a location to discover and download packages. Cargo is configured to use it by default to find requested packages. The project uses several dependencies, the most important of which include:

nom [6] A parser combinators library for Rust. Its goal is to provide tools to build safe parsers without compromising the speed or memory consumption. To that end, it uses extensively Rust's strong typing and memory safety to produce fast and correct parsers, and provides functions, macros and traits to abstract most of the error prone details.

While programming language parsers are usually written manually for more flexibility and performance, nom can be (and has been successfully) used as a prototyping parser for a language. The resulting code is small, and looks like the grammar the developer would have written with other parser approaches. The resulting parsers are small and easy to write, as well as easy to test separately. [**TODO** *[7]*]

unicode-segmentation [35] A library with a set of iterators which split strings on *grapheme clusters*, *words* or *sentence boundaries*, according to the Unicode Standard Annex #29 [8] rules.

wasm-bindgen [26] A Rust library and CLI tool that facilitate high-level interactions between Wasm modules and JavaScript. More specifically, this library allows JavaScript and Wasm to communicate with strings, JS objects, classes, etc, as opposed to purely integers and floats. Notable features of this project include:

- Importing JS functionality in to Rust such as DOM manipulation, console logging, or performance monitoring.
- Working with rich types like strings, numbers, classes, closures, and objects.

- Automatically generating TypeScript bindings for Rust code being consumed by JS.

Wasm-bindgen only generates bindings and glue for the JavaScript imports that are actually being used and Rust functionality that is being exported.

All of the above dependencies are available under the MIT license.

Svelte

Svelte [32] is a free and open-source front end JavaScript framework. Svelte has its own compiler for converting app code into client-side JavaScript at build time. The developer writes the components using HTML, CSS and JavaScript and during the build process Svelte compiles them into small standalone JavaScript modules. While frameworks like React and Vue do the bulk of their work in the user's browser while the app is running, Svelte shifts that work into a compile step that happens only when the developer builds their app, producing highly-optimized vanilla JavaScript. By statically analysing the component template, the compiler can make sure that the browser does as little work as possible. The outcome of this approach is not only smaller application bundles and better performance, but also a developer experience that is more approachable for people that have limited experience of the modern tooling ecosystem. Svelte is particularly appropriate to tackle the following situations:

- Web applications intended for low power devices: Applications built with Svelte have smaller bundle sizes, which is ideal for devices with slow network connections and limited processing power. Less code means less KB to download, parse, execute, and keep hanging around in memory.

- Highly interactive pages or complex visualizations: If you are building data-visualizations that need to display a large number of DOM elements, the performance gains that come from a framework with no runtime overhead will ensure that user interactions are snappy and responsive.

- Onboarding people with basic web development knowledge: Svelte has a shallow learning curve. Web developers with basic HTML, CSS, and JavaScript knowledge can easily grasp Svelte specifics in a short time and start building web applications.

Being a compiler, Svelte can extend HTML, CSS, and JavaScript, generating optimal JavaScript code without any runtime overhead. To achieve this, Svelte extends vanilla web technologies and only intervenes in very specific situations and only in the context of Svelte components.

Rollup

[TODO]

npm

[TODO]

5.2. Business logic

5.2.1. Grammar definition

[**TODO** *opis*]

5.2.2. Lexical analyser

[**TODO** *krótko o “algorytmie” tokenizacji*]

5.2.3. Syntactic analyser

[**TODO** *zdefiniowanie ważnych parserów dla EBNF*]

5.2.4. Left recursion handling

[**TODO** *przedstawienie algorytmu do usuwania lewej rekurencji i wyjaśnienie po co*]

5.2.5. Dependency graph reduction

[**TODO** *przedstawienie algorytmu do wyszukania reguły początkowej*]

5.2.6. Grammar processing

[**TODO** *opisanie sposobu na sprawdzenie czy wejście należy do języka generowanego przez gramatykę*]

5.3. Command line application

[**TODO**]

5.4. Web-based application

5.4.1. Linking the business logic

[**TODO** *jak się kompiluje Rusta do WebAssembly, czyli wasm-pack*]

5.4.2. Text editor

[**TODO** *CodeMirror*]

5.4.3. Visualizations

[**TODO**]

6. Project quality study

6.1. Business logic testing

6.1.1. Unit testing

[**TODO** *cargo test*]

6.1.2. Integration testing

[**TODO**]

6.2. UI testing

[**TODO** *Jest*]

6.3. Benchmarking

[**TODO** *cargo bench*]

6.4. Auditing

[**TODO** *Google Lighthouse*]

6.5. Complexity analysis

[**TODO** *clippy*]
[**TODO** *liczba linii kodu*]

7. Deployment

7.1. GitHub Pages

[TODO]

7.2. Electron

[TODO]

8. Artifacts

8.1. Source code

[TODO]

8.2. Web application

[TODO]

8.3. Desktop application

[TODO]

8.4. Command-line tool

[TODO]

8.5. Documentation

[TODO]

9. User manual

9.1. System requirements

[TODO]

9.2. Installation guide

[TODO]

9.3. Usage guide

[TODO]

10. Summary

[TODO]

Bibliography

[1] Aho, A., et al. *Kompilatory: reguły, metody, narzędzia*. Wydawnictwo Naukowe PWN, Warszawa, 2019, ch. 3,4.

[2] Aho, A. V. Algorithms for finding patterns in strings. In *Algorithms and Complexity*. Elsevier, 1990, pp. 255–300.

[3] Beazley, D. PLY homepage. `https://www.dabeaz.com/ply/`. Retrieved 24.10.2020.

[4] Chacon, S., and Straub, B. *Pro Git*. Apress, Berkeley, CA New York, NY, 2014.

[5] Chomsky, N. Three models for the description of language. *IEEE Transactions on Information Theory 2*, 3 (Sept. 1956), 113–124.

[6] Couprie, G. Nom GitHub page. `https://github.com/Geal/nom`. Retrieved 08.11.2020.

[7] Couprie, G. Nom, a byte oriented, streaming, zero copy, parser combinators library in Rust. In *2015 IEEE Security and Privacy Workshops* (May 2015), IEEE.

[8] Davis, M., and Chapman, C. Unicode standard annex #29. `http://www.unicode.org/reports/tr29/`. Retrieved 08.11.2020.

[9] Dib, F. Regex101 homepage. `https://regex101.com/`. Retrieved 24.10.2020.

[10] Driessen, V. A successful Git branching model. `https://nvie.com/posts/a-successful-git-branching-model/`, 2010. Retrieved 07.11.2020.

[11] Ferrous Systems. rust-analyzer homepage. `https://rust-analyzer.github.io/`. Retrieved 07.11.2020.

[12] Fokker, J. Functional parsers. In *Advanced Functional Programming*. Springer Berlin Heidelberg, 1995, pp. 1–23.

[13] Ford, B. Parsing expression grammars. In *Proceedings of the 31st ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL '04* (2004), ACM Press.

[14] Free Software Foundation. GNU Bison homepage. `https://www.gnu.org/software/bison/`. Retrieved 24.10.2020.

[15] Git community. Git homepage. `https://git-scm.com/`. Retrieved 07.11.2020.

[16] GitHub Inc. GitHub homepage. `https://github.com/`. Retrieved 08.11.2020.

[17] Hopcroft, J., et al. *Wprowadzenie do teorii automatów, języków i obliczeń*. Wydawnictwo Naukowe PWN, Warszawa, 2005, ch. 5.

[18] ISO/IEC. *ISO/IEC 14977:1996(E) — Information technology, syntactic metalanguage, Extended BNF*. Geneva, 1996.

[19] Johnson, W. L., Porter, J. H., Ackley, S. I., and Ross, D. T. Automatic generation of efficient lexical processors using finite state techniques. *Communications of the ACM 11*, 12 (Dec. 1968), 805–813.

[20] Klabnik, S., and Nichols, C. *The Rust programming language*. No Starch Press, Inc, San Francisco, 2018.

[21] Leijen, D., and Meijer, E. Parsec: Direct style monadic parser combinators for the real world.

[22] Meduna, A. *Formal Languages and Computation: Models and Their Applications*. CRC Press, Taylor & Francis Group, Boca Raton, 2014.

[23] Microsoft. Visual Studio Code homepage. `https://code.visualstudio.com/`. Retrieved 07.11.2020.

[24] Mössenböck, H., and Löberbauer, M. The compiler generator Coco/R homepage. `http://www.ssw.uni-linz.ac.at/Coco/`. Retrieved 24.10.2020.

[25] Parr, T. ANTLR homepage. `https://www.antlr.org/`. Retrieved 24.10.2020.

[26] Rust and WebAssembly team. Wasm-bindgen GitHub page. `https://github.com/rustwasm/wasm-bindgen`. Retrieved 08.11.2020.

[27] Rust Team. Rust homepage. `https://www.rust-lang.org/`. Retrieved 08.11.2020.

[28] Sipser, M. *Wprowadzenie do teorii obliczeń*. Wydawnictwa Naukowo-Techniczne, Warszawa, 2009, ch. 2.

[29] Stack Overflow. Developer Survey Results 2018. `https://insights.stackoverflow.com/survey/2018`. Retrieved 07.11.2020.

[30] Stack Overflow. Developer Survey Results 2019. `https://insights.stackoverflow.com/survey/2019`. Retrieved 07.11.2020.

[31] Svelte team. Svelte API documentation. `https://svelte.dev/docs`. Retrieved 24.10.2020.

[32] Svelte team. Svelte homepage. `https://svelte.dev`. Retrieved 08.11.2020.

[33] Svelte team. Svelte Language Tools GitHub page. `https://github.com/sveltejs/language-tools`. Retrieved 07.11.2020.

[34] Swierstra, S. D. Combinator parsing: A short tutorial. In *Language Engineering and Rigorous Software Development*. Springer Berlin Heidelberg, 2009, pp. 252–300.

[35] Unicode-rs team. Unicode-segmentation GitHub page. `https://github.com/unicode-rs/unicode-segmentation`. Retrieved 08.11.2020.

[36] WebAssembly team. WebAssembly homepage. `https://webassembly.org/`. Retrieved 08.11.2020.

List of Figures

List of Tables

List of Listings

A. Modified specification

[TODO]

```
character
  = ? any Unicode non-control character ?;
letter
  = ? any Unicode alphabetic character ?;
digit
  = ? any Unicode numeric character ?;
whitespace
  = ? any Unicode whitespace character ?;
comment
  = '(*', {comment | character}, '*)';
gap
  = (whitespace | comment), {whitespace}, {{comment}, {whitespace}};
identifier
  = letter, {{whitespace}, letter | digit};
factor
  = [[gap], digit, {{whitespace}, digit}, [gap], '*'],
    [gap], [(identifier
      | ('[' | '(/'), alternative, (']' | '/)')
      | ('{' | '(:'), alternative, ('}' | ':)')
      | '(', alternative, ')'
      | "'", character - "'", {character - "'"}, "'"
      | '"', character - '"', {character - '"'}, '"'
      | '?', {{whitespace}, character - '?'}, '?'), [gap]];
term
  = factor,
    ['-', ? a factor that could be replaced
      by a factor containing no identifiers ?];
sequence
  = term, {',', term};
alternative
  = sequence, {('|' | '/' | '!'), sequence};
production
  = [gap], identifier, [gap], '=', alternative, (';' | '.'), [gap];
grammar
  = production, {production};
```

Listing A.1: Modified version of the EBNF language specification defined in [18]