



**INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA**
Bahia

Campus
Irecê

Átila Gama Silva

Witch Cooking: Formatação de Código via Linguagem de Consulta do Tree-Sitter

Irecê
2023

Átila Gama Silva

Witch Cooking: Formatação de Código via Linguagem de Consulta do Tree-Sitter

Trabalho de Conclusão de Curso apresentado ao
*Curso Técnico em Análise e Desenvolvimento de Sis-
temas do Instituto Federal de Educação, Ciência e
Tecnologia da Bahia – Campus Irecê*, como requisito
parcial para obtenção do diploma de *Técnico em
Análise e Desenvolvimento de Sistemas*, mediante
a orientação do professor Rafael Xavier.

Irecê
2023

Resumo

A formatação de código é fundamental no desenvolvimento de software, permitindo estabelecer aspectos desejados como a padronização e legibilidade do código, que impactam positivamente o ciclo de vida do software. Em ambientes de desenvolvimento modernos, é comum o uso de ferramentas para automatizar a formatação de código. No entanto, essas ferramentas geralmente apresentam limitações no número de linguagens suportadas, nas opções de configuração e não permitem a definição de estilos de formatação personalizados. Visando superar ou reduzir as limitações frequentemente encontradas nas ferramentas convencionais de formatação, neste trabalho é apresentado o *Witch Cooking*, um software prototípico que tem como objetivos (I) abranger uma gama de linguagens de programação, além de (II) permitir que o usuário defina seus próprios estilos de formatação.

Palavras-chave: formatação de código; *prettyprint*; formatação personalizada; *Tree-Sitter*; linguagem de consulta.

Lista de Ilustrações

Figura 1 – Fluxograma do Algoritmo de Formatação	24
--	----

Lista de Trechos de Código

Trecho de Código 1 – Função <i>add_two</i> em C	22
Trecho de Código 2 – Função <i>add_two</i> em Rust	22
Trecho de Código 3 – Árvore Sintática da Função <i>add_two</i> em C	22
Trecho de Código 4 – Árvore Sintática da Função <i>add_two</i> em Rust	22
Trecho de Código 5 – Consulta Correspondendo a um Parâmetro de Função em C . .	22
Trecho de Código 6 – Consulta Correspondendo a um Parâmetro de Função em Rust .	22
Trecho de Código 7 – Consulta Correspondendo ao Parâmetro <i>x</i> de uma Função em C	22

Sumário

1	Introdução	11
2	A Formatação de Código-Fonte	13
3	O Sistema <i>Tree-Sitter</i>	17
4	O <i>Witch Cooking</i>	19
4.1	Motivações e Propósitos	19
4.2	Usagem	19
4.3	Materiais	19
4.4	Métodos	20
4.5	Fazendo Consultas	21
4.6	O Algoritmo de Formatação	23
5	Conclusão	25
	Referências	27
	Glossário	31

1 Introdução

A flexibilidade presente na sintaxe de linguagens de programação permite que diferentes arranjos dum mesmo código compartilhem um valor sintático equivalente. Essa característica possibilita a formatação do código de acordo com aspectos desejados, como a legibilidade, que é fundamental no ciclo de vida do software (BUSE; WEIMER, 2009, p. 546; OLIVEIRA et al., 2020, p. 1). Consequentemente, a formatação é frequentemente utilizada para estabelecer um nível satisfatório de legibilidade em bases de código.

A formatação de código pode ser realizada manualmente pelo programador, embora esse processo possa ser demorado, especialmente em grandes bases de código, exigindo do programador um tempo que poderia ser empregado em outra tarefa. Além disso, a formatação manual também pode ser falha e inconsistente devido à suscetibilidade humana ao erro, podendo ser agravada quando há múltiplos programadores em uma base de código.

Para evitar os problemas inerentes da formatação manual, é comum a utilização de softwares que automatizam a formatação de forma determinística, tornando o código padronizado e consistente. No entanto, é comum que essas ferramentas de formatação sejam limitadas a uma linguagem ou a uma família de linguagens de programação. Além disso, algumas delas são opinativas, assim restringindo as possibilidades de personalização pelo usuário. Finalmente, é também comum que essas ferramentas de formatação não permitam que o usuário defina estilos de formatação personalizados.

Em ambientes de desenvolvimento modernos, além da utilização de ferramentas que automatizam a formatação de código, é também comum a utilização do *Tree-Sitter* (TREE-SITTER..., 2023), um sistema de análise sintática de código aberto que foi disponibilizado ao público geral no GitHub primeiramente em 2019, tendo sido inicialmente desenvolvido por Max Brunsfeld. Desde seu lançamento, o *Tree-Sitter* tem ganhado popularidade na comunidade de desenvolvedores devido dentre outros motivos à (I) sua capacidade de suportar várias linguagens de programação; além de (II) sua linguagem de consulta, a qual permite a realização de buscas complexas na árvore analisada de um código-fonte.

Visando superar ou reduzir as limitações previamente mencionadas, quais são frequentemente presentes nas ferramentas convencionais de formatação, este trabalho tem como objetivo geral desenvolver um software prototípico para a formatação de código. Nomeada de *Witch Cooking* (SILVA, 2023b), a ferramenta desenvolvida tem como objetivos (I) abranger uma gama de linguagens de programação, além de (II) permitir que o usuário defina seus estilos personalizados de formatação.

Para alcançar o objetivo geral proposto, este trabalho tem como objetivos específicos (I) apresentar e contextualizar ferramentas de formatação de código conceituadas, além de abordar seus métodos de formatação; (II) conceituar, contextualizar e abordar as tecnologias do sistema *Tree-Sitter*; e, por fim, (III) utilizar o *Tree-Sitter* para desenvolver e atender às aspirações do *Witch Cooking*.

2 A Formatação de Código-Fonte

Desde os primórdios da computação, métodos foram desenvolvidos para garantir que a saída impressa fosse formatada de maneira esteticamente agradável (HARRIS, 1956 apud YELLAND, 2015b, p. 1). Esses métodos ganharam popularidade sob o termo “*prettyprinting*”, que se refere à formatação visual de diversos tipos de conteúdo. No desenvolvimento de software, o *prettyprinting* é conhecido como formatação de código, uma prática histórica e comum que envolve a adoção de convenções estilísticas para estruturar o código-fonte. Existem diversas terminologias utilizadas para se referir às ferramentas que realizam a formatação do código-fonte, incluindo (I) “formatadores de código”, ou “*code formatters*” em inglês; (II) “*prettyprinters*”; ou ainda (III) “*beautifiers*”.

Durante as décadas de 60 e 70, a linguagem de programação LISP¹ proporcionou condições favoráveis para o avanço da formatação de código (YELLAND, 2015a, p. 2). (I) LISP apresentava uma sintaxe distinta e expressiva, baseada em listas e estruturas aninhadas delimitadas por parênteses², o que tornava o código-fonte praticamente ilegível e incompreensível caso não fosse devidamente organizado. Além disso, (II) LISP era *homoicônico*, ou seja, permitia a manipulação do código via dados da própria linguagem. Essas duas características intrínsecas do LISP fundamentaram o surgimento de (I) *prettyprinters* para a linguagem, visando aperfeiçoar a legibilidade do código-fonte escrito ou emitido; além de (II) novas técnicas e abordagens para a formatação de código em geral. Esses avanços contribuíram significativamente para o conhecimento e o aprimoramento de práticas de formatação de código em trabalhos subsequentes.

Em 1967, Bill Gosper desenvolveu o GRINDEF (acrônimo para “*GRIND*³ *Function*”), considerado o primeiro *prettyprinter* a mensurar o tamanho das linhas e ter ciência de sua localização no arquivo (GOSPER, 2023; GRIESEMER, 2022). Essa ferramenta implementava o algoritmo *recursive re-predictor*, que, como descrito por Goldstein (1973), percorria a árvore de nós representando as listas e imprimia os nós considerando seus tamanhos e a quantidade máxima de caracteres que ainda poderiam ser inseridos em uma linha.

Posteriormente, Hearn e Norman (1979) propuseram um algoritmo mais elaborado para *prettyprinting* que utilizava um par de corrotinas⁴, (I) uma responsável por produzir uma sequência de caracteres que representavam o programa sendo impresso; e (II) outra responsável

¹ O termo “LISP” (acrônimo para “*LISt Processing*”) inicialmente se referia à linguagem de programação desenvolvida por McCarthy (1960), porém, com a disseminação de dialetos da linguagem, o termo também passou a ser utilizado para se referir à família de linguagens derivadas da original.

² Na notação original de McCarthy, eram utilizadas *M-expressions* entre colchetes para representar expressões. Essas *M-expressions* seriam posteriormente traduzidas em *S-expressions*. Por exemplo, a *M-expression* `car[cons[A; B]]` seria equivalente à *S-expression* `(car (cons A B))`. No entanto, assim que a linguagem LISP foi implementada, os programadores prontamente adotaram o uso das *S-expressions* em vez das *M-expressions*, tornando as *S-expressions* a forma predominante de representar a estrutura do código na linguagem.

³ O termo “*grind*” era utilizado em alguns círculos de LISP como sinônimo para *prettyprinters*.

⁴ Na obra, os autores enfatizam que a implementação com o uso de corrotinas é tão simples que poderia ser simulada – sem grandes dificuldades – em linguagens que não possuíam suporte nativo para corrotinas (HEARN; NORMAN, 1979, p. 53).

por decidir como esses caracteres seriam exibidos. Essas corrotinas se comunicavam por meio de um buffer **FIFO**, permitindo que as decisões de formatação fossem adiadas até que houvesse informações suficientes disponíveis para tomá-las com confiabilidade.

No ano seguinte, **Oppen (1980)** apresentou em sua influente obra, intitulada “*Prettyprinting*”, um algoritmo inovador de formatação de código-fonte. Embora apresentasse semelhanças ao algoritmo proposto por **Hearn e Norman (1979)**, destacava-se por sua capacidade de formatar código derivado de qualquer linguagem de programação. Para realizar essa tarefa, o algoritmo necessitava que o código-fonte fosse anotado com espaços em branco⁵ e delimitadores especiais para marcar o início e fim de blocos logicamente contíguos. Assim, o código a ser fornecido ao algoritmo precisaria ser processado por uma ferramenta intermediária capaz de compreender a sintaxe da linguagem e fornecer um código anotado de forma adequada, permitindo que o algoritmo realizasse a formatação apropriada.

Recentemente, tem se tornado cada vez mais evidente a crescente relevância dos softwares opinativos (**ECCLES, 2015**), também conhecidos em inglês como *opinionated softwares*. Esses softwares preestabelecem práticas estritivas, visando deliberadamente frustrar ou dificultar práticas destoantes (**LANCE, 2021**). Similarmente, os *prettyprinters* primitivos, tais como o SOAP (**SCOWEN et al., 1971**), eram naturalmente restritivos devido à simplicidade e às limitações dos algoritmos de formatação existentes. Conforme mencionado, à medida que os algoritmos de formatação de código-fonte eram aprimorados, tornando-se mais robustos e sofisticados, os *prettyprinters* passaram a proporcionar mais personalizações aos usuários. No entanto, atualmente, a abordagem opinativa também tem sido adotada por softwares modernos de formatação de código-fonte (**BLACK. . ., 2023**; **GRIESEMER, 2022**, p. 5–6, 8). Esses softwares geralmente têm como objetivos (I) padronizar e (II) garantir a consistência do código; além (III) eliminar o tempo e esforços necessários para escolher os estilos ideais de formatação; e, finalmente, consequentemente e principalmente (IV) aprimorar o ambiente de desenvolvimento colaborativo.

Um exemplo proeminente de *prettyprinter* opinativo é o **Black** (**BLACK. . ., 2023**), um formatador de código Python, cuja logomarca parodia a da fabricante automobilística Ford, e cujo irônico slogan “*any color you like*” faz alusão à famosa frase “*any customer can have a car painted any colour that he wants so long as it is black*” (**FORD; CROWTHER, 1922**, p. 72, grifo nosso), dita por Henry Ford, o fundador da fabricante. Assim como Ford adotou uma abordagem de produção em massa na fabricação dos *Model Ts* com uma única opção de cor para aumentar a eficiência (**FORD; CROWTHER, 1922**), o **Black** segue um princípio análogo. Ao impor um estilo estrito de formatação, minimizando o espaço para personalização ou preferências individuais, o **Black** oferece velocidade, determinismo, além de economia de tempo e esforços para questões mais importantes (**BLACK. . ., 2023**).

Sumarizando, desenvolver um *prettyprinter* não é uma tarefa simples (**HUGHES, 1995**, p. 55). Durante o processo de formatação, é necessário (I) ter conhecimento das estruturas gramaticais da linguagem específica em questão. Além disso, o objetivo do *prettyprinter* é (II)

⁵ Quebras de linha, avanços de formulário e alimentações de linha também eram tratados como espaços em branco pelo algoritmo.

formatar o código de maneira otimizada e agradável para o leitor (HUGHES, 1995, p. 55), levando em consideração as melhores práticas de formatação. Em alguns casos, também é importante (III) permitir que o usuário possa optar por estilos de formatação que atenda às suas preferências individuais. Adicionalmente, as sintaxes das linguagens de programação costumam ser flexíveis em relação à estruturação do código-fonte. Assim, para atender o item II, geralmente é necessário (IV) utilizar algoritmos que calculem o layout mais adequado de acordo com o contexto (JASPER, 2023, p. 12; GOLDSTEIN, 1973; HUGHES, 1995; YELLAND, 2015a). Fatalmente, essas características essenciais de um *prettyprinter* normalmente o restringem a uma linguagem específica ou, em alguns casos, a uma família de linguagens que compartilham semelhanças sintáticas ou estruturais.

3 O Sistema *Tree-Sitter*

O *Tree-Sitter* é um sistema multilíngue de análise sintática para ferramentas de programação inicialmente desenvolvido como um projeto secundário por Max Brunsfeld. Como relatado pelo próprio autor ([TREE-SITTER... , 2017](#)), o *Tree-Sitter* surgiu como uma tentativa de solucionar problemas presentes nas ferramentas de análise sintática da época. Mais especificamente, o sistema tinha como objetivos (I) ser utilizado no ambiente de desenvolvimento para produzir árvores de sintaxe a partir da análise de códigos escritos em várias linguagens; (II) implementar a análise incremental, permitindo a atualização da árvore de sintaxe em tempo real; (III) expor através da árvore de sintaxe os nós representando suas construções gramaticais no código (e.g., classes, funções, declarações, etc.), diferentemente das ferramentas contemporâneas, que utilizavam uma abordagem simplística baseada em expressões regulares; e, por fim, (IV) ser livre de dependências, assim beneficiando sua adoção e aplicabilidade.

Além das funcionalidades previamente mencionadas presentes no *Tree-Sitter*, o sistema também conta uma ferramenta de linha de comando que pode ser utilizada para gerar *parsers* para uma linguagem a partir de sua gramática. A gramática é definida via a linguagem de programação JavaScript, a qual (I) foi eleita 15 vezes seguidas pela *Stack Overflow Developer Survey* ([STACK... , 2022](#)) como a linguagem de programação mais comumente usada; além de (II) ser amplamente considerada pela comunidade de programadores como uma das linguagens mais fáceis de aprender e programar ([11... , 2023](#); [GOEL, 2023](#); [JAVASCRIPT... , 2023](#)). A ferramenta de geração de *parsers* também disponibiliza funções preestabelecidas para permitir a criação de gramáticas com diferentes níveis de complexidade. Não é surpreendente que, devido a essas características e facilidades presentes na criação de *parsers*, exista uma variedade de linguagens de programação e formatos de arquivos – variando de linguagens com sintaxes complexas, como C++ e Perl, a formatos de arquivos mais específicos, como *.lhs* e *.rasi* –, os quais têm *parsers* gerados pelo *Tree-Sitter* e, conseqüentemente, são suportados pelo sistema.

Na análise de código, é comum realizar tarefas que envolvem a busca de padrões na árvore sintática. Para isso, o *Tree-Sitter* oferece uma pequena linguagem de consulta declarativa que é capaz de expressar esses padrões por meio de *S-expressions* e buscar correspondências. A linguagem de consulta suporta operadores que permitem (I) a captura de nós; (II) a quantificação de nós, análoga às expressões regulares; (III) o agrupamento de nós; (IV) as alternâncias de nós; (V) o uso de *wildcards*; e (VI) a ordenação de nós. Adicionalmente, é permitido o uso de predicados – funções arbitrárias geralmente utilizadas para filtrar nós ou realizar verificações mais complexas durante a busca de padrões – sejam eles *builtins* ou estendidos por meio de uma API.

Um exemplo da relevância e utilidade da linguagem de consulta é o plugin *nvim-treesitter* ([NVIM-TREESITTER... , 2023](#)), que é frequentemente utilizado no editor *Neovim* ([NEOVIM... , 2023](#)). Esse plugin utiliza a linguagem de consulta para definir diferentes recursos, tais como (I) *code folding*, que permite ocultar blocos de código; (II) *highlights*, que realçam a sintaxe

do código; (III) indentações, que definem a estrutura do código; (IV) injeções, que permitem adicionar novas sintaxes a arquivos existentes; além de (V) captura de nós correspondentes a construções gramaticais (e.g., funções, classes, métodos, etc.), os quais são frequentemente utilizados em rotinas de programação tais como a remoção e navegação de código.

Em resumo, o sistema *Tree-Sitter* se mostra uma solução inovadora e eficiente para a análise sintática de códigos em diversas linguagens de programação, com uma abordagem diferenciada e sofisticada que possibilita a atualização em tempo real da árvore sintática e a identificação precisa das construções gramaticais presentes no código. Além disso, a ferramenta de linha de comando disponível no sistema facilita a geração de parsers a partir de gramáticas definidas em JavaScript, o que torna o processo mais acessível e personalizável para os programadores. Finalmente, a linguagem de consulta declarativa oferecida pelo *Tree-Sitter* se mostra uma importante *feature*, sendo utilizada em diversos plugins de editores de código para realizar tarefas variadas e sofisticadas, contribuindo significativamente no ambiente de desenvolvimento.

4 O Witch Cooking

4.1 Motivações e Propósitos

As motivações para o desenvolvimento do *Witch Cooking* surgiram das dificuldades enfrentadas pelo autor ao estudar estilos de formatação de código-fonte em diversas linguagens de programação com as quais ele trabalhava. O estudo envolvia a exploração e análise tanto de estilos convencionais¹, dos quais muitos eram amplamente adotados, quanto de estilos não-convencionais. Durante a análise dos estilos convencionais de formatação, (I) era imprescindível recorrer a diferentes *prettyprinters*, (II) cada um com suas próprias configurações e níveis de suporte para esses estilos. Por outro lado, durante a análise dos estilos não-convencionais, (III) a aplicação manual era inevitável, consumindo consideravelmente tempo e esforço.

Tendo em vista (I) as dificuldades mencionadas anteriormente; além da (II) complexidade envolvida no *prettyprinting* – (III) agravada pela proposta de formatação multilíngue –, o *Witch Cooking* foi concebido como um software de linha de comando – de natureza prototípica – para a formatação de código-fonte. A ferramenta elaborada tem como objetivos (I) abranger uma gama de linguagens de programação e (II) permitir que o usuário defina seus próprios estilos de formatação de código-fonte.

4.2 Usagem

O *Witch Cooking* apresenta a seguinte usagem `cook [-l LANG] -q QUERY [SRC]`, onde (I) `cook` é o nome do executável, (II) `[-l LANG]` é uma opção para explicitar a linguagem do código-fonte a ser formatado, (III) `-q QUERY` é utilizado para definir o caminho para o arquivo – escrito na linguagem de consulta do *Tree-Sitter* – no qual as instruções de formatação estão definidas, e (IV) `[SRC]` é um argumento opcional para especificar o caminho para o arquivo a ser formatado. Quando `[SRC]` é provido e `[-l LANG]` é omitido, o software tentará inferir uma linguagem conforme a extensão do nome do arquivo. Por outro lado, quando `[SRC]` é omitido, o código-fonte a ser formatado deve ser provido via *pipeline*, e a sua linguagem deve ser explicitada via `[-l LANG]`. Em alinhamento com sua natureza experimental, o *Witch Cooking* proporciona o código-fonte formatado resultante como saída, sem modificar o conteúdo original, efetivamente estabelecendo uma formatação segura.

4.3 Materiais

O *Witch Cooking* foi programado em Rust ([RUST...](#), 2023), que é uma linguagem amplamente reconhecida na comunidade de programadores ([STACK...](#), 2023). Dentre os motivos que levaram à escolha dessa linguagem para o desenvolvimento, destacam-se (I) seu gerencia-

¹ Neste trabalho, são considerados convencionais os estilos de formatação abordados nas páginas *Indentation style* e *Programming style* da Wikipédia. ([INDENTATION...](#), 2023; [PROGRAMMING...](#), 2023)

mento de memória seguro² (SOFTWARE... , 2022, p. 3); (II) seu excelente desempenho, a par às linguagens C e C++; (III) seus recursos de linguagens modernas, incluindo (A) expressividade, (B) *zero-cost abstractions*, (C) correspondência de padrões, (D) *closures*, (E) macros, etc.; (IV) seu suporte a múltiplas plataformas; além de (V) seu ecossistema de desenvolvimento fértil, que oferece (A) uma ampla variedade de bibliotecas, (B) *Cargo*, o gerenciador de pacotes, além do (C) *rust-analyzer* (RUST-ANALYZER... , 2023), uma implementação do *Language Server Protocol* (LSP) para a linguagem.

Para atender a seus objetivos, o *Witch Cooking* foi desenvolvido com base na biblioteca *Tree-Sitter* (TREE-SITTER... , 2023), a qual (I) suporta uma ampla variedade de linguagens de programação e formatos de arquivos, além de (II) possuir uma linguagem de consulta versátil e extensível que, dentre outras características, permite (A) a realização de buscas complexas na árvore de sintaxe analisada de um código-fonte, bem como (B) a adição de predicados personalizados.

Para escrever o código do *Witch Cooking*, foi utilizado o editor de texto *Neovim* (NEOVIM... , 2023), com configurações personalizadas (SILVA, 2023a). Essas configurações possibilitaram a integração com o *rust-analyzer*, o que resultou em melhorias significativas na experiência de programação. Dentre os plugins utilizados na configuração, destacam-se (I) o *nvim-treesitter* (NVIM-TREESITTER... , 2023), que permitiu a integração com o *Tree-Sitter*; além do (II) *playground* (PLAYGROUND... , 2023), que disponibilizou ferramentas para uma melhor compreensão do *Tree-Sitter* e de sua linguagem de consulta.

4.4 Métodos

Embora o *Tree-Sitter* não tenha sido originalmente projetado para formatação de código-fonte, o autor deste trabalho identificou o potencial da biblioteca, particularmente em relação à sua linguagem de consulta, que – como já ressaltado – é extensível. Assim, para permitir a formatação instruída, o *Witch Cooking* baseou-se na extensão e disponibilização de predicados designados a executar operações específicas de formatação. Em outras palavras, a concepção do autor era que a formatação ocorresse conforme a aplicação dos predicados estendidos em padrões especificados no arquivo de instruções, escrito na linguagem de consulta do *Tree-Sitter*.

Consequentemente, no desenvolvimento deste trabalho, foram utilizados diferentes procedimentos, incluindo (I) pesquisa bibliográfica e (II) pesquisa documental, com o intuito de aprofundar o conhecimento sobre o funcionamento do *Tree-Sitter*; além de (III) pesquisa experimental, assim como (IV) estudo de caso, com as respectivas finalidades de (A) explorar a aplicação do *Tree-Sitter* e sua linguagem de consulta no processo de formatação de código-fonte, além de (B) analisar a eficácia do algoritmo de formatação desenvolvido. Essa abordagem metodológica se caracteriza como predominantemente qualitativa, de natureza básica e orientada por objetivos exploratórios.

² Embora Rust não apresente *garbage collection* – o que contribui significativamente para o desempenho –, a linguagem emprega *features* e conceitos como *borrow checker*, *ownership* e *lifetime*, garantindo um gerenciamento de memória seguro. Isso é especialmente relevante em linguagens de alto desempenho que não possuem coleta de lixo (e.g., C e C++). Estudos estimam que, nesse contexto, pelo menos 65% das vulnerabilidades de segurança sejam causadas por falta de segurança no gerenciamento de memória (GAYNOR, 2020).

4.5 Fazendo Consultas

Como mencionado na seção 4.2, o *Witch Cooking* formata o código-fonte conforme as instruções presentes no arquivo submetido via `-q QUERY`, o qual é escrito na linguagem de consulta do *Tree-Sitter*. Devido à importância fundamental desempenhada por essa linguagem na ferramenta, é essencial possuir um conhecimento básico sobre o seu funcionamento. Portanto, nesta seção, será abordado superficialmente o funcionamento da linguagem de consulta do *Tree-Sitter*, permitindo assim uma melhor compreensão da base em que o *Witch Cooking* foi desenvolvido.

Como abordado no capítulo 3, a linguagem de consulta do *Tree-Sitter* foi originalmente concebida para buscar padrões na árvore sintática via (I) definição de padrões de correspondência e (II) aplicação de predicados. Os predicados disponíveis nativamente pelo *Tree-Sitter* são agnósticos quanto à linguagem de programação utilizada. Por outro lado, os padrões de correspondência variam de acordo com a linguagem devido às suas construções gramaticais específicas, que são posteriormente traduzidas como nós na árvore de sintaxe analisada.

Por exemplo, considere os trechos de código 1 e 2, escritos nas linguagens de programação C e Rust, respectivamente. Embora ambos desempenhem papéis idênticos – i.e., definem uma função que retorna a soma do parâmetro x com 2 –, suas árvores de sintaxe analisadas – apresentadas respectivamente nos trechos de código 3 e 4 – exibem diferenças significativas. Portanto, para buscar a correspondência do trecho em que o parâmetro x é definido em ambos os códigos (i.e., `int add_two(int x) {` e `fn add_two(x: i32) -> i32 {`), são necessárias consultas distintas, apresentadas pelos trechos de código 5 e 6.

Dado que os padrões correspondem a qualquer uma de suas ocorrências, é importante garantir maior especificidade em sua definição. Isso pode ser obtido por meio de (I) especificação de campos, (II) especificação de nós anônimos e (III) aplicação de predicados. Para ilustrar, ao aplicar a consulta apresentada no trecho de código 5 ao seguinte código em linha `int add(int x, int y);`, obtêm-se duas correspondências, como destacadas no seguinte trecho `int add(int x, int y);`. Logo, para corresponder somente ao parâmetro x , deve-se utilizar uma consulta mais elaborada, conforme exemplificada no trecho de código 7. Nesse trecho, é possível notar algumas diferenças em relação à sua versão original, que foi apresentada no trecho de código 5. (I) O nó *identifier* é capturado com o nome *id*, como destacado no seguinte trecho de código `declarator: (identifier) @id`. Além disso, (II) o predicado *eq?* é aplicado para garantir que o conteúdo da captura *id* corresponda à string “x” o que, por consequência, impede a correspondência do parâmetro y .

Por fim, é importante ressaltar que a linguagem de consulta do *Tree-Sitter* disponibiliza recursos adicionais (TREE-SITTER..., 2023) que vão além do escopo desta seção. No entanto, as informações abordadas até então estabelecem uma base fundamental para compreender o funcionamento da linguagem de consulta e sua aplicação em conjunto com o *Witch Cooking* na formatação do código-fonte.

Trecho de Código 1 – Função *add_two* em C

```

1 int add_two(int x) {
2     return x + 2;
3 }

```

Fonte autoral.

Trecho de Código 2 – Função *add_two* em Rust

```

1 fn add_two(x: i32) -> i32 {
2     x + 2
3 }

```

Fonte autoral.

Trecho de Código 3 – Árvore Sintática da Função *add_two* em C

```

1 function_definition
2   type: primitive_type
3   declarator: function_declarator
4     declarator: identifier
5     parameters: parameter_list
6       parameter_declaration
7         type: primitive_type
8         declarator: identifier
9   body: compound_statement
10  return_statement
11  binary_expression
12    left: identifier
13    right: number_literal

```

Fonte autoral. Nós anônimos foram omitidos.

Trecho de Código 4 – Árvore Sintática da Função *add_two* em Rust

```

1 function_item
2   name: identifier
3   parameters: parameters
4     parameter
5       pattern: identifier
6       type: primitive_type
7   return_type: primitive_type
8   body: block
9     binary_expression
10      left: identifier
11      right: integer_literal

```

Fonte autoral. Nós anônimos foram omitidos.

Trecho de Código 5 – Consulta Correspondendo a um Parâmetro de Função em C

```

1 (parameter_declaration
2   type: (primitive_type)
3   declarator: (identifier))

```

Fonte autoral.

Trecho de Código 6 – Consulta Correspondendo a um Parâmetro de Função em Rust

```

1 (parameter
2   pattern: (identifier)
3   type: (primitive_type))

```

Fonte autoral.

Trecho de Código 7 – Consulta Correspondendo ao Parâmetro *x* de uma Função em C

```

1 (parameter_declaration
2   type: (primitive_type)
3   declarator: (identifier) @id
4   (#eq? @id "x"))

```

Fonte autoral.

4.6 O Algoritmo de Formatação

O *Witch Cooking* permite a formatação de código-fonte via seu algoritmo fundamentado no *Tree-Sitter* e sua linguagem de consulta. Esse algoritmo utiliza a biblioteca para gerar uma árvore de sintaxe a partir do código-fonte submetido pelo usuário. Adicionalmente, complementando os predicados *builtins* do *Tree-Sitter* – os quais foram concebidos para a busca de padrões correspondentes –, o algoritmo inclui predicados desenvolvidos para executar procedimentos variados de formatação. Consequentemente, ao utilizar esses predicados nos padrões da consulta, é possível instruir a formatação de nós específicos da árvore de sintaxe.

A operação do algoritmo de formatação divide-se nas etapas de (I) inicialização, (II) formatação, e (III) retorno, como ilustrado na fig. 1. Durante a etapa de inicialização – onde estruturas são instanciadas –, primeiramente, (I) a árvore de sintaxe – derivada do código-fonte submetido para formatação – é analisada. Em seguida, (II) um objeto³ de consulta é gerado a partir do arquivo contendo as instruções de formatação, escrito na linguagem de consulta do *Tree-Sitter*. Esse objeto (A) contém informações sobre padrões, capturas, propriedades e predicados; e (B) é posteriormente utilizado para obter as correspondências dos seus padrões.

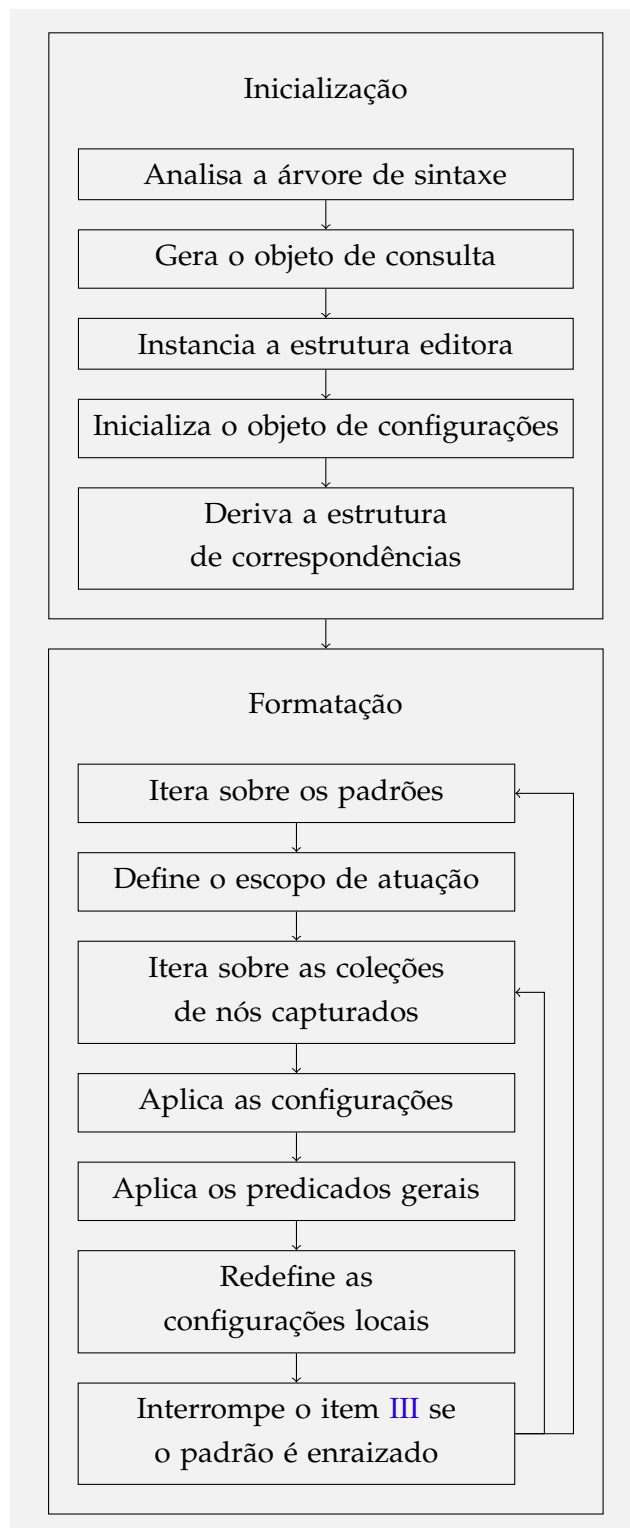
Além disso, (III) é instanciada uma estrutura editora que contém (A) o texto do código-fonte em edição – originado do código submetido pelo usuário –, juntamente com (B) um vetor de edições. Essa estrutura é utilizada pelos predicados para realizar as edições no código e armazená-las. Na sequência, (IV) é inicializado um objeto de configurações, que é futuramente empregado pelos predicados para obter e definir configurações em diferentes níveis de escopo, sejam eles globais – i.e., abrangendo toda a consulta – ou locais – i.e., restrito a um padrão. Finalizando a etapa de inicialização, (V) a estrutura de correspondências é derivada do (A) objeto de consulta, juntamente com (B) a árvore analisada e (C) o texto do código-fonte submetido. Simplificadamente, essa estrutura pode ser descrita como um iterador que fornece nós que correspondem a capturas presentes em padrões específicos.

O processo efetivo de formatação de código-fonte inicia na segunda etapa. Nessa fase, a dinâmica se baseia em um *loop* que (I) itera sobre todos os padrões especificados no arquivo de instruções de formatação. Em cada iteração, inicialmente, (II) é checado se o padrão é enraizado⁴ ou não, determinando assim se o escopo de atuação do padrão é local ou global, respectivamente. Em seguida, (III) um segundo *loop* percorre as coleções de nós capturados para cada correspondência do padrão em questão. No corpo dessa estrutura de repetição, (IV) são aplicadas as configurações especificadas no padrão via o predicado *builtin set!*. Além disso, (V) são aplicados os predicados gerais – i.e., predicados não-*builtins* – especificados no padrão em iteração. Pressupõe-se que esses predicados estejam dentre os providos pelo *Witch Cooking* para instruir a formatação.

³ Neste contexto, o termo “objeto” é utilizado como uma alternativa para descrever uma estrutura em Rust.

⁴ Na linguagem de consulta do *Tree-Sitter*, embora a biblioteca não forneça uma definição exata, um padrão enraizado é geralmente caracterizado por especificar pelo menos um nó inicial para a correspondência, e vice-versa.

Figura 1 – Fluxograma do Algoritmo de Formatação



Fonte autoral.

5 Conclusão

Referências

- 11 Most In-Demand Programming Languages. Berkeley Extension. Disponível em: <<https://bootcamp.berkeley.edu/blog/most-in-demand-programming-languages>>. Acesso em: 5 mai. 2023.
- BLACK: The uncompromising Python code formatter. Versão 23.3.0. Black. Disponível em: <<https://black.readthedocs.io/>>. Acesso em: 8 mai. 2023.
- BUSE, Raymond P. L.; WEIMER, Westley R. Learning a Metric for Code Readability. **IEEE Transactions on Software Engineering**, IEEE, v. 36, n. 4, p. 546–558, 2009. DOI: [10.1109/TSE.2009.70](https://doi.org/10.1109/TSE.2009.70).
- ECCLES, Stuart. **The Rise of Opinionated Software**. Medium. 28 mai. 2015. Disponível em: <<https://medium.com/@stueccles/the-rise-of-opinionated-software-calba0140d5b>>. Acesso em: 26 mai. 2023.
- FORD, Henry; CROWTHER, Samuel. **My Life and Work**. Edição: Samuel Crowther. [S.l.]: Doubleday, Page & Company, 1922. ISBN 1594621985.
- FRIED, Jason; HANSSON, Heinemeier David; LINDERMAN, Matthew. **Getting Real**: The smarter, faster, easier way to build a successful web application. [S.l.]: 37signals, 18 nov. 2009. ISBN 0578012812. Disponível em: <<https://basecamp.com/gettingreal>>. Acesso em: 26 mai. 2023.
- GAYNOR, Alex. **What science can tell us about C and C++'s security**. 27 mai. 2020. Disponível em: <<https://alexgaynor.net/2020/may/27/science-on-memory-unsafety-and-security/>>. Acesso em: 26 set. 2023.
- TREE-SITTER: a new parsing system for programming tools - GitHub Universe 2017. 1 vídeo (43 min). GitHub. 2017. Disponível em: <<https://youtu.be/a1rC79DHpmY>>. Acesso em: 30 abr. 2023.
- GOEL, Aman. **How to Learn JavaScript in 2023 | 8 Best Ways For Beginners**. Disponível em: <<https://hackr.io/blog/how-to-learn-javascript>>. Acesso em: 4 mai. 2023.
- GOLDSTEIN, Ira. **Pretty printing**: Converting list to linear structure. [S.l.], fev. 1973.
- GOSPER, Ralph William. **Twubblesome Twelve**. Disponível em: <<http://gospers.org/bill.html>>. Acesso em: 21 mai. 2023.
- GRIESEMER, Robert. **The Cultural Evolution of gofmt**. Google Research. 2022. Disponível em: <<https://go.dev/talks/2015/gofmt-en.slide>>. Acesso em: 20 mai. 2023.
- HARRIS, R. W. Keyboard Standardization. **Western Union Technical Review**, v. 10, n. 1, p. 37–42, 1956.
- HEARN, Anthony C.; NORMAN, Arthur C. A one-pass prettyprinter. **ACM SIGPLAN Notices**, v. 14, n. 12, p. 50–58, 1979. DOI: [10.1145/954004.954005](https://doi.org/10.1145/954004.954005).

- HUGHES, John. The Design of a Pretty-printing Library. In: **Advanced Functional Programming: First International Spring School on Advanced Functional Programming Techniques**, Båstad, Sweden, May 1995, Tutorial Text. Edição: Johan Jeuring e Erik Meijer. Båstad, Sweden: Springer, 15 mai. 1995. v. 925, p. 53–96. (Lecture Notes in Computer Science). DOI: [10.1007/3-540-59451-5_3](https://doi.org/10.1007/3-540-59451-5_3).
- JASPER, Daniel. **clang-format**: Automatic formatting for C++. Disponível em: <http://lvm.org/devmtg/2013-04/jasper-slides.pdf>. Acesso em: 27 mai. 2023.
- LANCE, Corrina. **About opinionated software**: guidance VS freedom. Medium. 27 jul. 2021. Disponível em: <https://medium.com/codex/about-opinionated-software-guidance-vs-freedom-2bd66de304fe>. Acesso em: 27 mai. 2023.
- MCCARTHY, John. Recursive Functions of Symbolic Expressions and Their Computation by Machine, Part I. **Communications of the ACM**, v. 3, n. 4, p. 184–195, 1 abr. 1960. DOI: [10.1145/367177.367199](https://doi.org/10.1145/367177.367199).
- NATIONAL SECURITY AGENCY. **Software Memory Safety**: Cybersecurity Information Sheet. [S.l.], 22 nov. 2022. Disponível em: https://media.defense.gov/2022/Nov/10/2003112742/-1/-1/0/CSI_SOFTWARE_MEMORY_SAFETY.PDF. Acesso em: 29 set. 2023.
- NEOVIM: hyperextensible Vim-based text editor. Versão 0.9.0. Neovim. Disponível em: <https://neovim.io/>. Acesso em: 7 mai. 2023.
- NVIM-TREESITTER: Nvim Treesitter configurations and abstraction layer. Versão 0.9.0. nvim-treesitter. Disponível em: <https://github.com/nvim-treesitter/nvim-treesitter>. Acesso em: 7 mai. 2023.
- PLAYGROUND: Treesitter playground integrated into Neovim. nvim-treesitter. Disponível em: <https://github.com/nvim-treesitter/playground>. Acesso em: 2 out. 2023.
- OLIVEIRA, Delano and et al. Evaluating Code Readability and Legibility: An Examination of Human-centric Studies. In: IEEE. 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME). [S.l.: s.n.], 2020. DOI: [10.1109/icsme46990.2020.00041](https://doi.org/10.1109/icsme46990.2020.00041).
- OPPEN, Derek C. Prettyprinting. **ACM Transactions on Programming Languages and Systems**, v. 2, n. 4, p. 465–483, out. 1980. DOI: [10.1145/357114.357115](https://doi.org/10.1145/357114.357115).
- SCOWEN, Roger S. et al. SOAP: A Program which Documents and Edits ALGOL 60 Programs. **The Computer Journal**, v. 14, n. 2, p. 133–135, 1971. DOI: [10.1093/comjnl/14.2.133](https://doi.org/10.1093/comjnl/14.2.133).
- SILVA, Átila Gama. **UMA DLÇ**: A monkey-flavored configuration soup for Neovim. Disponível em: <https://github.com/atchim/uma-dlc>. Acesso em: 1 out. 2023.
- _____. **Witch Cooking**: Experimental multilingual code formatter based on Tree-Sitter’s query. Disponível em: <https://github.com/atchim/witch-cooking>. Acesso em: 27 abr. 2023.

- STACK Overflow Developer Survey. Stack Overflow. 2022. Disponível em:
<<https://survey.stackoverflow.co/2022>>. Acesso em: 1 mai. 2023.
- STACK Overflow Developer Survey. Stack Overflow. 2023. Disponível em:
<<https://survey.stackoverflow.co/2023/#section-admired-and-desired-programming-scripting-and-markup-languages>>. Acesso em: 26 set. 2023.
- RUST: A language empowering everyone to build reliable and efficient software. Versão 1.72.0. The Rust Programming Language. Disponível em: <<https://www.rust-lang.org/>>. Acesso em: 18 set. 2023.
- RUST-ANALYZER: Bringing a great IDE experience to the Rust programming language. Versão 2023-09-25. The Rust Programming Language. Disponível em:
<<https://rust-analyzer.github.io/>>. Acesso em: 1 out. 2023.
- TREE-SITTER: a parsing system for programming tools. Versão 0.20.8. Tree-Sitter. Disponível em: <<https://tree-sitter.github.io/>>. Acesso em: 6 abr. 2023.
- TREE-SITTER: Using Parsers. Tree-Sitter. Disponível em:
<<https://tree-sitter.github.io/tree-sitter/using-parsers#pattern-matching-with-queries>>. Acesso em: 22 set. 2023.
- JAVASCRIPT Tutorial. W3Schools. Disponível em: <<https://www.w3schools.com/js>>. Acesso em: 5 mai. 2023.
- INDENTATION style. Wikipedia. Disponível em:
<https://en.wikipedia.org/wiki/Indentation_style>. Acesso em: 29 mai. 2023.
- PROGRAMMING style. Wikipedia. Disponível em:
<https://en.wikipedia.org/wiki/Programming_style>. Acesso em: 29 mai. 2023.
- YELLAND, Phillip M. A New Approach to Optimal Code Formatting. **Google Research**, 2015. Disponível em: <<https://research.google.com/pubs/archive/44667.pdf>>. Acesso em: 7 abr. 2023.
- _____. **rfmt**: R source code formatter. Versão 1.0. 5 ago. 2015. Disponível em:
<<https://github.com/google/rfmt>>. Acesso em: 23 mai. 2023.

Glossário

B

Builtin No desenvolvimento de software, “*builtin*” é um termo utilizado para descrever funções ou comandos que são incorporados no próprio sistema ou ambiente de programação, em contraste com funções ou comandos definidos pelo usuário ou programador. Por exemplo, em Python, a função *print* é um exemplo de função *builtin*. Essas funções ou comandos *builtin* estão prontamente disponíveis, o que significa que não é necessário definir ou importá-los explicitamente por parte do usuário ou programador.

C

Code folding No desenvolvimento de software, *code folding* é uma *feature* que permite a ocultação seletiva de partes de um documento ou código-fonte, permitindo ao usuário ou programador visualizar apenas as seções de interesse. Esse recurso é especialmente útil ao lidar com grandes volumes de informações, possibilitando uma melhor organização e foco em seções específicas. O *code folding* é comumente utilizado em ambientes de programação que possuem estruturas hierárquicas, como árvores ou blocos aninhados, facilitando a navegação e a compreensão do código.

F

Feature Na programação de computadores, uma *feature* é uma funcionalidade específica ou capacidade de um software que agrega valor ao sistema. Uma *feature* pode ser um componente, uma função, um conjunto de comandos ou qualquer outra característica que ofereça uma determinada funcionalidade aos usuários do software. Ela pode abranger desde recursos básicos até funcionalidades mais avançadas e personalizadas, e é projetada para atender às necessidades e demandas dos usuários, melhorar a usabilidade do software e fornecer uma experiência mais completa.

FIFO Na computação e na teoria de sistemas, “FIFO” é um acrônimo para “*First In, First Out*” – o primeiro a entrar é o primeiro a sair –, um método para organizar a manipulação de uma estrutura de dados – geralmente, especificamente um buffer de dados – onde a entrada mais antiga (primeira), ou “cabeça” da fila, é processada primeiro.

G

Garbage collection (GC) Em ciência da computação, *garbage collection* (GC), ou coleta de lixo, é uma forma de gerenciamento automático de memória. O coletor de lixo tenta

recuperar a memória que foi alocada pelo programa, mas que não está mais sendo referenciada; essa memória é chamada de lixo. A coleta de lixo foi inventada pelo cientista da computação americano John McCarthy por volta de 1959 para simplificar o gerenciamento manual de memória em Lisp.

H

Homoicônico Na programação de computadores, a homoiconicidade – das palavras gregas “*homo-*”, que significa “o mesmo”, e “*icon*”, que significa “representação” – é uma propriedade de algumas linguagens de programação. Uma linguagem é homoicônica se um programa escrito nela pode ser manipulado como dados utilizando a própria linguagem, permitindo inferir a representação interna do programa apenas lendo o próprio programa. Essa propriedade é frequentemente resumida dizendo que a linguagem trata o **código como dados**.

L

Loop Na programação de computadores, *loop* – também conhecido em português como “estrutura de repetição” ou “laço de repetição” – é uma sequência de declarações que é especificada uma vez, mas que pode ser executada várias vezes em sucessão. O código “dentro” do *loop* – o corpo do *loop* – é executado (I) um número especificado de vezes, ou (II) uma vez para cada item de uma coleção, ou (III) até que alguma condição seja satisfeita, ou (IV) indefinidamente.

O

Opinionated No desenvolvimento de software, “*opinionated*” é um termo utilizado para se referir a um conjunto de práticas preestabelecidas sobre como abordar uma determinada tarefa, podendo não permitir customizações ou desvio das abordagens. O termo foi popularizado a partir da obra *Getting Real: The smarter, faster, easier way to build a successful web application*, escrita por [Fried, Hansson e Linderman \(2009\)](#).

P

Parser Na programação de computadores, um *parser* (ou analisador) é um componente de software que recebe dados de entrada (frequentemente texto) e constrói uma estrutura de dados – frequentemente algum tipo de árvore de análise, árvore de sintaxe abstrata ou outra estrutura hierárquica – fornecendo uma representação estrutural da entrada enquanto verifica a sintaxe correta. A análise pode ser precedida ou seguida por outras etapas, ou essas etapas podem ser combinadas em uma única etapa. O *parser* é frequentemente precedido por um analisador léxico (*lexer*) separado, que cria tokens a partir da sequência de caracteres de entrada; alternativamente, esses elementos podem ser combinados na análise sem scanner. *Parsers*

podem ser programados manualmente ou podem ser gerados automaticamente ou semi-automaticamente por um gerador de *parser*.

Pipeline Em sistemas operacionais de computador semelhantes ao Unix, *pipeline* é um mecanismo para comunicação entre processos usando passagem de mensagens. Um *pipeline* é um conjunto de processos encadeados por seus fluxos padrão, de modo que o texto de saída de cada processo (*stdout*) é passado diretamente como entrada (*stdin*) para o próximo. O segundo processo é iniciado enquanto o primeiro processo ainda está executando, e eles são executados simultaneamente.

Prettyprinting (PP) Na programação de computadores, *prettyprinting* é a aplicação de diversas convenções estilísticas de formatação a arquivos de texto, como código-fonte, marcação e conteúdos similares. Essas convenções de formatação podem incluir o uso de estilos de indentação, cores e tipos de fonte diferentes para destacar elementos sintáticos do código-fonte, ou ajustes de tamanho, para tornar o conteúdo mais fácil de ser lido e compreendido por pessoas. *Prettyprinters* para código-fonte são às vezes chamados de “formatadores de código” ou “*beautifiers*”.

S

S-expression (Sexp) Na programação de computadores, uma *S-expression* – ou expressão simbólica, abreviada como “*sexpr*” ou “*sexp*” – é uma expressão em uma notação de mesmo nome para dados em lista aninhada (estruturados em árvore). As *S-expressions* foram inventadas e popularizadas pela linguagem de programação LISP, que as utiliza tanto para código-fonte quanto para dados.

W

Wildcard No software, um *wildcard* é um tipo de espaço reservado geralmente representado por um único caractere – como um asterisco (“*”) – ou símbolo especial, que pode representar (I) qualquer outro caractere; um (II) conjunto de caracteres; (III) objetos, elementos, símbolos, etc.; ou até (IV) nada, ou seja, nenhum caractere ou objeto, elemento, símbolo, etc. Os usos mais comuns de *wildcards* são em (I) expressões regulares (*regex*), onde um caractere especial é usado para representar um conjunto de caracteres desconhecidos ou variáveis em uma string; e na (II) interface de linha de comando, onde as *shells* normalmente disponibilizam o caractere asterisco (“*”) para corresponder a nomes de arquivos.

Z

Zero-cost abstractions Na programação de computadores, *Zero-cost abstractions* (ou abstrações de custo zero) é um conceito que se refere à capacidade de usar abstrações de alto nível e recursos da linguagem sem incorrer em penalidades de desempenho em tempo de execução. Isso significa que o programador pode escrever código de forma

mais abstrata, expressiva e amigável para o desenvolvedor, enquanto o sistema subjacente gera código de máquina altamente eficiente.