

به نام خدا

طراحی کامپایلر

آرش شفیعی



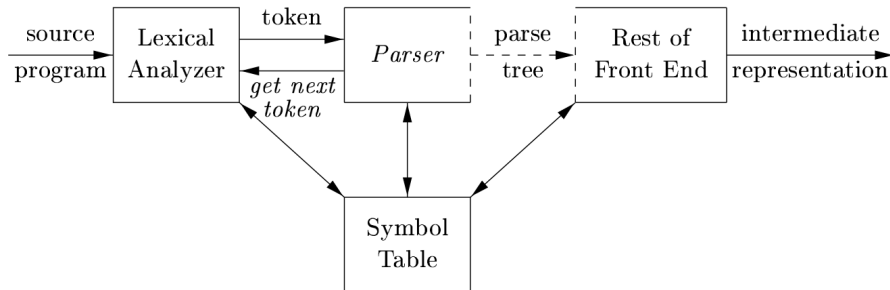
تحليل نحوي

- در این فصل در مورد الگوریتم‌های مختلف تجزیه گرامرها صحبت خواهیم کرد که معمولاً در کامپایلر استفاده می‌شوند.
- ساختار هر زبان برنامه نویسی توسط قوانینی تعیین می‌شود. برای مثال در زبان سی، یک برنامه از تعدادی توابع تشکیل شده است و هر تابع از تعدادی دستورات تشکیل شده که این دستورات می‌توانند تعریف و اعلام متغیرها، انتساب مقدار، دستورات شرطی و حلقه‌های تکرار باشند.
- ساختار نحوی¹ یک زبان، نحوه قرارگیری توکن‌ها در جملات زبان را تعیین می‌کند.
- ساختار نحوی یک زبان را می‌توان توسط گرامرهای مستقل از متن توصیف کرد.

¹ syntax

- توصیف ساختار نحوی توسط گرامر مستقل از متن دارای مزیت‌های زیر است:
- گرامرها می‌توانند توصیف بسیار دقیق و قابل فهمی از یک زبان برنامه‌نویسی ارائه می‌کنند.
- همچنین ابزارهایی وجود دارند که قادرند با دریافت گرامر یک زبان، به طور خودکار یک تجزیه کننده تولید کنند. استفاده از چنین ابزارهایی کمک می‌کنند که در صورتی که گرامر مشکلاتی داشته باشد، مشکلات آن به طور خودکار تشخیص داده شوند. برای مثال طراح یک گرامر ممکن است قادر به تشخیص ابهام در گرامر نباشد، درحالی که ابزار ممکن است ابهام‌ها را تشخیص دهد.
- یک طراح کامپایلر نیاز دارد برای طراحی درست تجزیه کننده توصیف دقیقی از زبان مورد نظر داشته باشد.
- وقتی یک کامپایلر براساس قوانین یک گرامر ساخته شده باشد، با تغییر گرامر به سادگی می‌توان برنامه تجزیه کننده کامپایلر را نیز تغییر داد.

- تحلیل‌گر نحوی¹ یا تجزیه‌کننده² (پارسر) توکن‌ها را از تحلیل‌گر لغوی دریافت می‌کند و بررسی می‌کند آیا دنباله توکن‌های دریافت شده می‌توانند توسط زبان گرامر توصیف شده تولید شوند یا خیر.



¹ syntax analyzer

² parser

- یک تجزیه کننده همچنین معمولاً قادر است هر نوع خطای نحوی را گزارش کرده و ادامه رسته ورودی را پس از خطا تجزیه کنند.
- یک تجزیه کننده با دریافت توکن‌ها، یک درخت تجزیه تولید می‌کند و درخت تجزیه تولید شده را به قسمت بعدی کامپایلر ارسال می‌کند.

- سه دسته از تجزیه کننده‌ها برای گرامرها وجود دارند : تجزیه کننده‌های عمومی¹ ، بالا به پایین² ، و پایین به بالا³ .
- الگوریتم‌های تجزیه عمومی مانند الگوریتم سی‌وای‌کا⁴ و الگوریتم ایرلی⁵ می‌توانند هر نوع الگوریتم مستقل از متن را تجزیه کنند. مشکل اصلی این تجزیه کننده‌ها این است که پیچیدگی زمانی بالایی دارند. گرچه پیچیدگی سی‌وای‌کا $O(n^3)$ و پیچیدگی الگوریتم ایرلی در بدترین حالت $O(n^3)$ است و از لحاظ تئوری پیچیدگی پایینی به حساب می‌آید ولی در عمل برای پیاده‌سازی کامپایلرها به تجزیه کننده‌هایی نیاز داریم که پیچیدگی زمانی پایین‌تری داشته باشند.

¹ universal

² top-down

³ bottom-up

⁴ Cocke-Younger-kasami (CYK)

⁵ Earley

- معمولاً در کامپایلرها از تجزیه کننده‌های بالا به پایین و پایین به بالا استفاده می‌شود.
- همانطور که از اسم این تجزیه کننده‌ها مشخص است، تجزیه کننده‌های بالا به پایین درخت تجزیه را از ریشه به برگ می‌سازند، درحالی که تجزیه کننده‌های پایین به بالا از برگ‌های درخت تجزیه آغاز می‌کنند تا به ریشه درخت برسند و درخت تجزیه را تشکیل دهند.
- در هر صورت ورودی تجزیه کننده دنباله‌ای از توکن‌هاست که از چپ به راست خوانده می‌شود.

- تجزیه کننده‌های بالا به پایین و پایین به بالا برای زیر مجموعه‌ای از گرامرهای مستقل از متن کارایی دارند، اما برخی از این گرامرهای خاص به خصوص گرامرهای LL و LR برای توصیف همه ساختارهای زبان‌های برنامه‌نویسی موجود کافی هستند.

- معمولاً بسیاری از ساختارهای زبان‌های برنامه‌نویسی پیچیدگی خاصی برای تجزیه ندارند. برای مثال یک حلقه `while` در زبان جاوا از کلمه `while`، یک عبارت درون یک جفت پرانتز و یک جفت آکولاد تشکیل شده است.
- عبارات ریاضی معمولاً به علت اولویت و وابستگی عملگرها پیچیدگی بیشتری دارند. بنابراین در اینجا بر روی عبارات ریاضی تمرکز می‌کنیم.
- با در نظر گرفتن تنها عملگرهای جمع، ضرب، و پرانتز، یک عبارت¹ به نام `E` تشکیل شده است از مجموع تعدادی جمله² به نام `T` که با عملگر `+` با یکدیگر جمع شده‌اند و هریک از جملات تشکیل شده است از ضرب تعدادی فاکتور (ضریب)³ به نام `F` که با استفاده از عملگر `*` در یکدیگر ضرب شده‌اند. هریک از فاکتورها می‌تواند یک شناسه باشد، و یا خود یک عبارت باشد که در بین دو پرانتز قرار گرفته است.

¹ expression

² term

³ factor

- بنابراین می‌توانیم گرامری به صورت زیر برای توصیف یک عبارت بنویسیم.

$$\begin{aligned} E &\rightarrow E + T \mid T \\ T &\rightarrow T * F \mid F \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$

- این گرامر به دسته گرامرهای LR تعلق دارد. این نوع گرامرها را معمولاً توسط تجزیه کننده پایین به بالا تجزیه می‌کنیم.

- این گرامر را نمی‌توانیم توسط تجزیه‌کننده بالا به پایین تجزیه کنیم زیرا بازگشتی چپ¹ است. در یک گرامر بازگشتی چپ قانونی وجود دارد که در آن متغیر سمت چپ بدنه قانون با متغیر سمت چپ قانون برابر است. خواهیم دید که تجزیه‌کننده بالا به پایین نمی‌تواند گرامرهایی که بازگشت چپ دارند را تجزیه کند.

¹ left recursive

- گرامر زیر معادل گرامر قبل و غیربازگشتی چپ¹ است و می‌توانیم از یک تجزیه کننده بالا به پایین برای تجزیه برنامه‌ها توسط آن استفاده کنیم. در مورد روش حذف بازگشت چپ توضیح خواهیم داد.

$$\begin{aligned}
 E &\rightarrow T E' \\
 E' &\rightarrow + T E' \mid \epsilon \\
 T &\rightarrow F T' \\
 T' &\rightarrow * F T' \mid \epsilon \\
 F &\rightarrow (E) \mid \text{id}
 \end{aligned}$$

- همچنین گرامر زیر یک گرامر مبهم است که برای رشته $a + b * c$ بیشتر از یک درخت تجزیه می‌سازد. گرامرهای مبهم را نیز قبل از تجزیه باید رفع ابهام کنیم.

$$E \rightarrow E + E \mid E * E \mid (E) \mid \text{id}$$

¹ non-left-recursive

- اگر قرار بود کامپایلرها فقط برنامه‌های درست را تجزیه کنند، طراحی و پیاده سازی آنها بسیار ساده‌تر می‌شد. اما، یک کامپایلر باید علاوه بر کامپایل برنامه به برنامه‌نویس کمک کند مکان و نوع خطاهای برنامه خود را شناسایی کند.

- خطاهای برنامه‌نویسی می‌توانند انواع مختلفی داشته باشند.
- ۱. خطاهای لغوی مانند خطا در نوشتن نام شناسه‌ها، کلمات کلیدی و غیره.
- ۲. خطاهای نحوی مانند خطا در نوشتن اشتباه ساختار دستورات.
- ۳. خطاهای معنایی مانند خطا در انتساب مقدار متغیرها با نوع متفاوت.
- ۴. خطاهای منطقی که شامل خطاهایی می‌شوند که در یک برنامه اتفاق می‌افتند هنگامی که برنامه از نظر لغوی و نحوی و معنایی درست است و برنامه به درستی کامپایل می‌شود اما نتیجه برنامه با مقدار مورد انتظار برنامه‌نویس متفاوت است. برای مثال در زبان سی ممکن است به اشتباه برنامه‌نویس به اشتباه به جای عملگر تساوی از عملگر انتساب استفاده کند.

- وقتی یک پارسر با خطا مواجه شد می‌تواند کامپایل را متوقف کند و اولین خطایی که با آن مواجه شده است را گزارش کند. اما بهتر است کامپایلر همه خطاهای یک برنامه را با یک بار تجزیه کد تشخیص دهد. برای این کار لازم است پس از مواجه شدن با یک خطا، تجزیه کننده خود را بازیابی کند و تجزیه برنامه را ادامه دهد.
- بنابراین کامپایلر باید علاوه بر تشخیص خطا و گزارش خطا به طور دقیق، بتواند سریعاً پس از رخداد یک خطا بازیابی شده و بررسی برنامه را ادامه دهد تا خطاهای بعدی را تشخیص دهد. همچنین مدیریت خطا نباید سربار زیادی بر روند کامپایل داشته باشد و باعث کندی بیش از اندازه کامپایل برنامه‌ها شود.

- چند استراتژی برای بازیابی از خطا وجود دارد که به آنها اشاره می‌کنیم.
- بازیابی با توکن همگام‌کننده یا بازیابی اضطراری¹ : در این روش تجزیه‌کننده از توکن‌ها یک‌به‌یک چشم پوشی می‌کند تا به یکی از توکن‌های همگام‌کننده² برسد. برای مثال علامت آکولاد بسته (}) یا نقطه ویرگول (;) می‌توانند توکن‌های همگام‌کننده باشند. مشکل این روش این است که ممکن است تعداد زیادی از خطاها نادیده گرفته شوند اما مزیت آن سادگی پیاده‌سازی آن است.
- بازیابی با جایگزینی توکن‌ها³ : با رخداد خطا، تجزیه‌کننده می‌تواند توکن‌های بعدی در ورودی را جایگزین کند تا جایی که ادامه رشته معنی‌دار و قابل تجزیه باشد. برای مثال با تبدیل یک علامت ویرگول به نقطه ویرگول ممکن است ورودی معنی‌دار و قابل تجزیه شود.

¹ panic-mode recovery

² synchronizing tokens

³ phrase-level recovery

- قوانین گرامری تشخیص خطا¹ : با پیش‌بینی کردن خطاهای معمول برنامه‌نویسی می‌توان تعدادی قوانین گرامری به گرامر اضافه کرد که خطاها را تشخیص می‌دهند.
- تصحیح عمومی و بهینه² : معمولاً انتظار داریم تجزیه‌کننده کمترین تعداد تصحیح را در ورودی انجام دهد. الگوریتم‌هایی وجود دارند که می‌توانند از بین چندین روش برای تصحیح گرامر، گزینه‌ای را انتخاب کنند که با استفاده از آن کمترین تصحیح بر روی ورودی صورت گیرد. این الگوریتم‌ها معمولاً بسیار پرهزینه هستند و معمولاً در عمل استفاده نمی‌شوند.

¹ error production rules

² global correction

گرامرهای مستقل از متن

- گرامرهای مستقل از متن می‌توانند ساختار نحوی زبان‌های برنامه‌نویسی را توصیف کنند. این گرامرها به ازای هریک از مفاهیم در زبان برنامه‌نویسی یک متغیر تعریف می‌کنند.
- برای مثال اگر مفاهیم دستور $^1 (stmt)$ و عبارت $^2 (expr)$ را در نظر بگیریم، می‌توانیم قانون گرامر زیر را تعریف کنیم.

`stmt` \rightarrow `if` (`expr`) `stmt` `else` `stmt`

- با استفاده از قوانین دیگر می‌توانیم تعریف کنیم یک دستور چه شکل‌های دیگری می‌تواند داشته باشد.

¹ statement

² expression

گرامرهای مستقل از متن

- یک گرامر مستقل از متن تشکیل شده است از نمادهای پایانی یا ترمینالها، نمادهای غیرپایانی یا متغیرها، یک نماد آغازین و تعدادی قوانین تولید.

۱. ترمینالها¹ یا نمادهای پایانی یا پایانهها واحدهایی هستند که رشته ورودی را تشکیل می‌دهند. ترمینالها در یک گرامر یک زبان برنامه‌نویسی همان توکن‌ها هستند. برای مثال کلمات کلیدی `if` و `else` و کاراکترهای (و) ترمینالهای یک گرامر هستند.

۲. نمادهای غیرپایانی² یا غیرپایانهها یا متغیرها دنباله‌ای از توکن‌ها را با یک نام انتزاعی نامگذاری می‌کنند. برای مثال متغیر `stmt` نماینده مفهوم دستور است که مقدار آن می‌تواند هریک از دستورات زبان باشد.

¹ terminal

² nonterminal

گرامرهای مستقل از متن

۳. یکی از متغیرها به عنوان نماد آغازین^۱ استفاده می‌شود.

۴. قوانین تولید^۲ یک گرامر تعیین می‌کنند چگونه متغیرها و ترمینال‌ها در کنار یکدیگر قرار می‌گیرند تا یک رشته از یک زبان را تشکیل دهند. هر قانون تولید تشکیل شده است از یک متغیر سمت چپ یا متغیر قانون تولید یا متغیر ابتدای قانون تولید^۳، یک نماد \rightarrow که گاهی با $=::$ نشان داده می‌شود و یک بدنه یا سمت راست^۴ قانون که از صفر یا چند ترمینال و متغیر تشکیل شده است. در فرایند تجزیه یک رشته با متغیر آغازین شروع می‌کنیم و متغیر را با بدنه یکی از قوانین تولید مربوط به آن جایگزین می‌کنیم. این فرایند را ادامه می‌دهیم تا رشته به دست بیاید. در صورتی که رشته مورد نظر به دست نیامد، رشته عضو گرامر آن زبان نیست. مجموعه همه رشته‌هایی که با شروع از نماد آغازین و اعمال قوانین یک گرامر به دست می‌آیند، زبان آن گرامر را تعیین می‌کنند.

^۱ start symbol

^۲ production rule

^۳ head or left side

^۴ body or right side

- برای مثال گرامر زیر عبارات ریاضی را تجزیه می‌کند که شامل عملگرهای + و - و * و / و (و) هستند. کلمه id درواقع نوع توکن شناسه است که در تحلیل لغوی استخراج شده است.

$expression \rightarrow expression + term$

$expression \rightarrow expression - term$

$expression \rightarrow term$

$term \rightarrow term * factor$

$term \rightarrow term / factor$

$term \rightarrow factor$

$factor \rightarrow (expression)$

$factor \rightarrow id$

گرامرهای مستقل از متن

- در این فصل از علائم و نشانه‌گذاری‌های زیر استفاده می‌کنیم.
- ترمینال‌ها شامل موارد زیر هستند : حروف کوچک ابتدایی الفبای انگلیسی مانند *a* و *b* و *c*، عملگرها مانند *+* و ***، علائم نشانه‌گذاری مانند پرانتز و کاما، ارقام مانند *0* و *1* و *۰۰۰* و *9* و رشته‌های پررنگ مانند *id* و *if*.
- متغیرها شامل موارد زیر هستند : حروف بزرگ ابتدایی الفبای انگلیسی مانند *A* و *B* و *C*، حرف *S* که بیشتر به عنوان متغیر آغازین استفاده می‌شود، رشته‌هایی که به صورت مورب نوشته می‌شود مانند *expr* و *stmt*.

گرامرهای مستقل از متن

- معمولاً وقتی می‌خواهیم از یک نماد گرامر، که ممکن است ترمینال یا متغیر باشد، صحبت کنیم آن را با حروف X و Y و Z نمایش می‌دهیم.
- یک رشته شامل ترمینال‌ها را معمولاً با حروف u و v و w و z نمایش می‌دهیم.
- برای نمایش دنباله‌ای از ترمینال‌ها و متغیرها از حروف یونانی مانند α و β استفاده می‌کنیم. مثلاً $A \rightarrow \alpha$ یک قانون گرامر است.
- وقتی یک متغیر چندین بدنه داشته باشد آنها را با علامت خط عمودی از یکدیگر جدا می‌کنیم مثلاً $A \rightarrow \alpha_1 | \alpha_2 | \dots | \alpha_k$
- معمولاً متغیر سمت چپ اولین قانون همان متغیر آغازین است.

- گرامر زیر عبارات ریاضی را توصیف می‌کند که در آن از متغیرهای E و T و F و ترمینال‌های +، -، *، /، (،) و id استفاده شده است.

$$\begin{aligned} E &\rightarrow E + T \mid E - T \mid T \\ T &\rightarrow T * F \mid T / F \mid F \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$

- به فرایندی که در آن یک رشته توسط قوانین یک گرامر تولید می‌شود، فرایند اشتقاق¹ گفته می‌شود.
- با شروع از نماد آغازین، در هرگام یکی از متغیرها با بدنه یکی از قوانین متعلق به آن متغیر جایگزین می‌شود. دنباله ترمینال‌ها و متغیرهایی که در هرگام به دست می‌آید را یک صورت جمله‌ای² می‌نامیم. اگر با جایگزین کردن متغیرها در صورت‌های جمله‌ای توسط بدنه قوانین متعلق به آنها، رشته مورد نظر به دست آمد، آن رشته متعلق به زبان گرامر است. در این صورت می‌گوییم رشته توسط گرامر مشتق می‌شود یا تولید می‌شود یا به دست می‌آید.

¹ derivation

² sentential form

- برای مثال، گرامر زیر با یک متغیر E را در نظر بگیرید.

$$E \rightarrow E + E \mid E * E \mid - E \mid (E) \mid \text{id}$$

- فرض کنید می‌خواهیم جمله (id) - توسط این گرامر به دست آوریم. می‌توانیم با اعمال سه قانون این رشته را به دست آوریم. می‌گوییم E با استفاده از قانون سوم مشتق می‌کند یا به دست می‌دهد $-E$ و سپس با استفاده از قانون چهارم به دست می‌دهد $-(E)$ و در نهایت با استفاده از قانون پنجم به دست می‌دهد $-(id)$.

$$E \Rightarrow -E \Rightarrow -(E) \Rightarrow -(id)$$

- به این دنباله از جایگزینی متغیرها یا بدنه قوانین متعلق به آنها یک فرایند اشتقاق می‌گوییم.

- دنباله‌ای از نمادها به صورت $\alpha A \beta$ را در نظر بگیرید به طوری که α و β دنباله‌ای از نمادهای پایانی و غیرپایانی (ترمینال‌ها و متغیرهای) گرامر هستند و A یک نماد غیرپایانی (متغیر) است.
- فرض کنید $A \rightarrow \gamma$ یک قانون تولید باشد. آنگاه می‌نویسیم $\alpha A \beta \Rightarrow \alpha \gamma \beta$. نماد \Rightarrow به معنی مشتق کردن در یک گام¹ است.

¹ derives in one step

- وقتی دنباله‌ای به صورت $\alpha_1 \Rightarrow \alpha_2 \Rightarrow \dots \Rightarrow \alpha_n$ داشته باشیم به طوری که $n \geq 1$ ، می‌گوییم α_1 از α_n مشتق می‌شود و یا α_1 به دست می‌دهد α_n و یا α_1 در صفر یا چند گام مشتق می‌کند α_n . به عبارت دیگر:

۱. به ازای هر صورت جمله‌ای α داریم $\alpha \xRightarrow{*} \alpha$ یعنی هر صورت جمله‌ای می‌تواند خود را در صفر یا چند گام^۱ مشتق کند.

۲. اگر $\alpha \xRightarrow{*} \beta$ و $\beta \Rightarrow \gamma$ آنگاه $\alpha \xRightarrow{*} \gamma$

- همچنین گاهی می‌نویسیم $\xRightarrow{+}$ به معنی مشتق کردن در یک یا چند گام.

- اگر $\alpha \xRightarrow{*} S$ جایی که S نماد آغازین گرامر G است، می‌گوییم α یک صورت جمله‌ای^۲ از گرامر G است.

- یک صورت جمله‌ای شامل متغیرها و ترمینال‌هاست. یک جمله^۳ از یک گرامر یک صورت جمله‌ای است که در آن هیچ متغیری نباشد.

^۱ derives in zero or more steps

^۲ sentential form

^۳ sentence

- زبان تولید شده توسط یک گرامر مجموعه‌ای است از همه جمل‌های تولید شده توسط آن گرامر.
- رشته w در زبان تولید شده توسط گرامر G یا $L(G)$ است اگر و تنها اگر w یک جمله از گرامر G باشد یا به عبارت دیگر $w \xRightarrow{*} S$.
- زبانی که توسط یک گرامر مستقل از متن تولید می‌شود، یک زبان مستقل از متن نام دارد.
- اگر دو گرامر، یک زبان یکسان تولید کنند، آن دو گرامر معادل یکدیگرند.

- گرامر زیر را در نظر بگیرید.

$$E \rightarrow E + E \mid E * E \mid - E \mid (E) \mid \mathbf{id}$$

- جمله $-(id + id)$ یک جمله از این گرامر است زیرا فرایند اشتقاق زیر برای آن وجود دارد :

$$E \Rightarrow -E \Rightarrow -(E) \Rightarrow -(E + E) \Rightarrow -(id + E) \Rightarrow -(id + id)$$

- می‌نویسیم $-(id + id) \xRightarrow{*} E$ و می‌خوانیم جمله $-(id + id)$ از متغیر E مشتق می‌شود.

- در هر گام در فرایند اشتقاق دو انتخاب وجود دارد. باید انتخاب کنیم کدام متغیر را جایگزین کنیم و همچنین کدام قانون متعلق به متغیر انتخاب شده را انتخاب کنیم.

- دو نوع فرایند اشتقاق را به صورت زیر تعریف می‌کنیم :

۱. در اشتقاق چپ^۱ ، متغیری که در صورت جمله‌ای در سمت چپ بقیه متغیرها قرار دارد و به عبارت دیگر چپ‌ترین^۲ است، انتخاب می‌شود. اگر $\alpha \Rightarrow \beta$ گامی باشد که در آن چپ‌ترین متغیر α انتخاب شود، می‌نویسیم $\alpha \xRightarrow{lm} \beta$.

۲. در اشتقاق راست^۳ ، متغیری که راست‌ترین^۴ است انتخاب می‌شود و می‌نویسیم $\alpha \xRightarrow{rm} \beta$.

^۱ leftmost derivation

^۲ leftmost

^۳ rightmost derivation

^۴ rightmost

- برای مثال :

$$E \xRightarrow{rm} -E \xRightarrow{rm} -(E) \xRightarrow{rm} -(E + E) \xRightarrow{rm} -(E + id) \xRightarrow{rm} -(id + id)$$

- به طور خلاصه می‌گوییم $wA\gamma \xRightarrow{lm} w\delta\gamma$ جایی که w فقط از ترمینال‌ها تشکیل شده و $A \rightarrow \delta$ یک قانون تولید است و γ رشته‌ای است تشکیل شده از متغیرها و ترمینال‌ها.

- اگر $S \xRightarrow{*}_{lm} \alpha$ آنگاه می‌گوییم α یک صورت جمله‌ای چپ¹ از گرامر است.

¹ left sentential form

- درخت تجزیه¹ یک نمایش گرافیکی از فرایند تجزیه است که در آن ترتیب جایگزینی متغیرها نشان داده نمی‌شود.
- هر رأس میانی در درخت تجزیه، اعمال یک قانون در فرایند اشتقاق را نشان می‌دهد. اگر یک رأس با برچسب A در درخت تجزیه داشته باشیم، فرزندان آن از سمت چپ به راست به ترتیب ترمینال‌ها و متغیرهایی هستند که در بدنه یکی از قوانین متعلق به A قرار دارند.

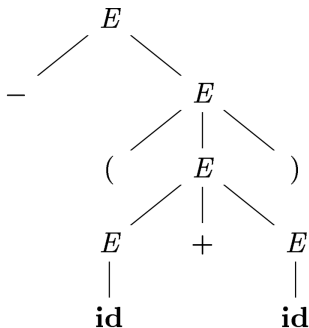
¹ parse tree

درخت تجزیه

- گرامر زیر را در نظر بگیرید.

$$E \rightarrow E + E \mid E * E \mid - E \mid (E) \mid \text{id}$$

- درخت تجزیه زیر برای به دست آوردن رشته $-(\text{id} + \text{id})$ با استفاده از این گرامر تشکیل شده است.



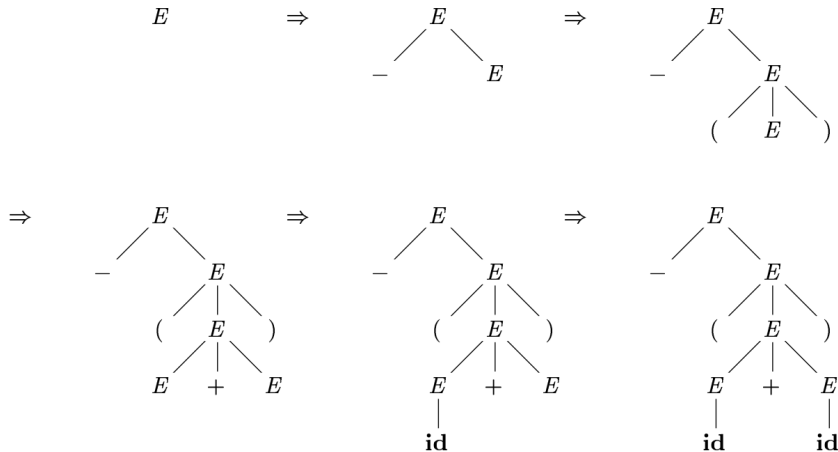
- برگ‌های درخت تجزیه همه با ترمینال‌ها برچسب زده شده‌اند و به ترتیب از چپ به راست رشته‌ای را تشکیل می‌دهند که توسط گرامر مشتق شده است.
- به رشته‌ای که از الحاق برگ‌های درخت تجزیه از چپ به راست به دست می‌آید محصول¹ درخت تجزیه گفته می‌شود.
- یک درخت تجزیه می‌تواند تجزیه یک صورت جمله‌ای را نشان دهد. به درخت تجزیه‌ای که محصول آن یک صورت جمله‌ای باشد، درخت تجزیه جزئی² نیز گفته می‌شود. به درخت تجزیه‌ای که محصول آن یک جمله باشد، درخت تجزیه کامل³ گفته می‌شود.

¹ yield

² partial parse tree

³ complete parse tree

- در شکل زیر درخت‌های تجزیه برای رشته $-(id + id)$ در فرایند اشتقاق چپ نشان داده شده‌اند.



- یک درخت تجزیه می‌تواند متناظر با چند فرایند اشتقاق باشد. مثلاً دو فرایند اشتقاق چپ و اشتقاق راست می‌توانند یک درخت تجزیه واحد تولید کنند.
- درخت تجزیه زیر می‌تواند توسط یک اشتقاق چپ یا یک اشتقاق راست تولید شده باشد.



- گرامری که بیش از یک درخت تجزیه برای یک جمله تولید کند، مبهم¹ نامیده می‌شود.
- به عبارت دیگر یک گرامر مبهم برای تولید یک رشته بیش از یک فرایند اشتقاق چپ (یا بیش از یک فرایند اشتقاق راست) دارد.

¹ ambiguous

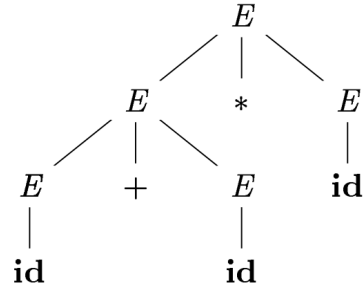
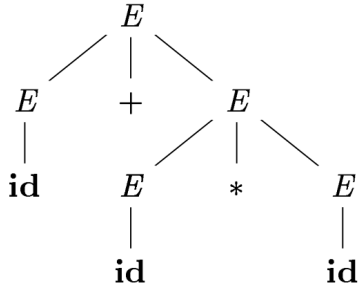
- گرامر زیر را در نظر بگیرید.

$$E \rightarrow E + E \mid E * E \mid - E \mid (E) \mid \text{id}$$

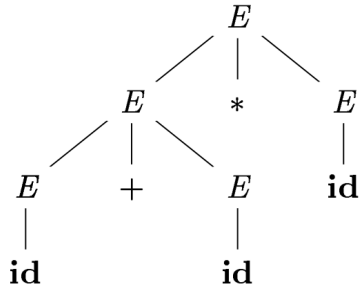
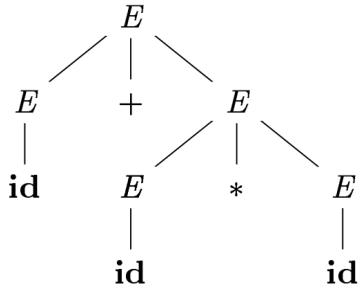
- برای به دست آوردن رشته $\text{id} + \text{id} * \text{id}$ توسط این گرامر دو فرایند اشتقاق چپ وجود دارد.

$ \begin{aligned} E &\Rightarrow E + E \\ &\Rightarrow \text{id} + E \\ &\Rightarrow \text{id} + E * E \\ &\Rightarrow \text{id} + \text{id} * E \\ &\Rightarrow \text{id} + \text{id} * \text{id} \end{aligned} $	$ \begin{aligned} E &\Rightarrow E * E \\ &\Rightarrow E + E * E \\ &\Rightarrow \text{id} + E * E \\ &\Rightarrow \text{id} + \text{id} * E \\ &\Rightarrow \text{id} + \text{id} * \text{id} \end{aligned} $
--	--

- همچنین برای این رشته دو درخت تجزیه به صورت زیر وجود دارد.



- دقت کنید که این دو درخت تجزیه دو معنی متفاوت از رشته تولید شده به دست می‌دهند. اگر بخواهیم رشته $a + b * c$ را توسط این گرامر تجزیه کنیم، درخت سمت چپ معادل $a + (b * c)$ و درخت سمت راست معادل $(a + b) * c$ خواهد بود.



گرامرهای منظم

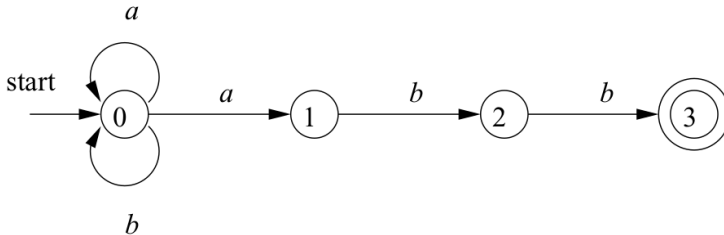
- گرامرها ابزار قوی‌تری نسبت به عبارات منظم هستند. هر عبارت منظم را می‌توان توسط یک گرامر نشان داد ولی هر گرامر را نمی‌توان توسط یک عبارت منظم نمایش داد. دسته‌ای از گرامرها که برای توصیف زبان‌های منظم به کار می‌روند، گرامرهای منظم نامیده می‌شوند. گرامرهای منظم زیر مجموعه‌ای از گرامرهای مستقل از متن هستند.

- عبارت منظم $(a|b)^*abb$ را می‌توان توسط گرامر منظم زیر توصیف کرد.

$$\begin{aligned}A_0 &\rightarrow aA_0 \mid bA_0 \mid aA_1 \\A_1 &\rightarrow bA_2 \\A_2 &\rightarrow bA_3 \\A_3 &\rightarrow \epsilon\end{aligned}$$

گرامرهای منظم

- الگوریتمی وجود دارد که توسط آن می‌توان یک ماشین متناهی غیرقطعی را به یک گرامر تبدیل کرد.
- گرامر قبل در واقع از ماشین متناهی غیرقطعی زیر به دست می‌آید.



- این الگوریتم به صورت زیر عمل می‌کند :

۱. به ازای هر حالت i از ماشین متناهی غیرقطعی متغیر A_i را می‌سازیم.
۲. اگر حالت i با ورودی a به حالت j می‌رود آنگاه قانون $A_i \rightarrow aA_j$ را به گرامر اضافه می‌کنیم. اگر حالت i با ورودی ϵ به حالت j می‌رود آنگاه قانون $A_i \rightarrow A_j$ را به گرامر اضافه می‌کنیم.
۳. اگر حالت i یک حالت نهایی است آنگاه قانون $A_i \rightarrow \epsilon$ را اضافه می‌کنیم.
۴. اگر حالت i یک حالت شروع است، آنگاه A_i را متغیر آغازین قرار می‌دهیم.

- برخی از زبان‌ها را نمی‌توانیم توسط یک گرامر منظم توصیف کنیم. این زبان‌ها متعلق به دسته زبان‌های منظم نیستند و ماشین متناهی برای آنها وجود ندارد.
- برای مثال $L = \{a^n b^n | n \geq 1\}$ زبانی است که نمی‌توان برای توصیف آن از یک ماشین متناهی استفاده کرد. توسط لم تزریق اثبات می‌شود که این زبان متعلق به دسته زبان‌های منظم نیست، اما می‌توان آن را توسط یک گرامر مستقل از متن توصیف کرد.

- گرامرهای مستقل از متن می‌توانند زبان‌های برنامه‌نویسی را توصیف کنند. البته گرامرها قادر به توصیف معنایی زبان‌ها نیستند. برای مثال توسط گرامر مستقل از متن نمی‌توانیم نیاز یک متغیر به تعریف قبل از استفاده از آن را توصیف کنیم.
- برای این که یک گرامر برای تجزیه‌کننده قابل استفاده باشد، باید پردازش‌هایی بر روی آن انجام شود که در اینجا به آنها اشاره می‌کنیم. برای مثال یک گرامر ابتدا باید رفع ابهام شود. سپس برای استفاده در تجزیه‌کننده بالا به پایین باید بازگشت چپ در آن حذف شود.

- همانطور که گفته شد، زبان‌های منظم را نیز می‌توان توسط گرامرها توصیف کرد. سؤالی که در اینجا ممکن است به وجود آید این است که چرا نیاز است که یک تحلیل‌گر لغوی قبل از تحلیل‌گر نحوی داشته باشیم؟
- با جدا کردن تحلیل‌گر لغوی از تحلیل‌گر نحوی تجزیه‌کننده بسیار ساده‌تر می‌شود و برنامه کامپایلر ساده‌تر می‌شود که باعث می‌شود تعداد خطاهای برنامه‌نویسی در نوشتن کامپایلر کاهش پیدا کند و همچنین برنامه کامپایلر ساده‌تر شود و راحت‌تر بتوان آن را تغییر داد. همچنین قوانین در تحلیل‌گر لغوی نسبتاً ساده‌اند و با عبارت‌های منظم ساده تولید می‌شوند و نیازی به افزودن گرامرهای پیچیده برای آنها وجود ندارد. به علاوه روشی وجود دارد که تحلیل‌گر لغوی مستقیماً از عبارت منظم تولید می‌شود. به این دلایل تحلیل‌گر لغوی از تحلیل‌گر نحوی جدا می‌شود.

- گاهی می‌توانیم یک گرامر غیر مبهم معادل یک گرامر مبهم بنویسیم. اما این کار همیشه ممکن نیست زیرا برخی از زبان‌ها ذاتاً مبهم هستند.
- گرامر مبهم زیر را در نظر بگیرید.

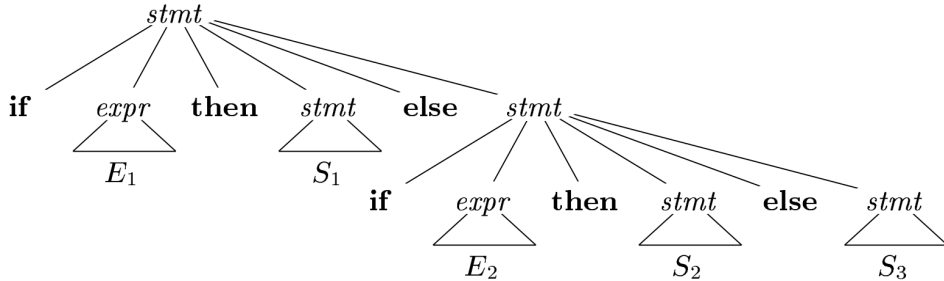
$$\begin{array}{lcl} stmt & \rightarrow & \text{if } expr \text{ then } stmt \\ & | & \text{if } expr \text{ then } stmt \text{ else } stmt \\ & | & \text{other} \end{array}$$

- در اینجا **other** به معنی هر دستور دیگری به غیر از دستورات شرطی if-else است.

- با استفاده از این گرامر می‌توانیم جمله زیر را تولید کنیم.

if E_1 then S_1 else if E_2 then S_2 else S_3

- برای این جمله درخت تجزیه زیر وجود دارد.



- حال جمله زیر را در نظر بگیرید.

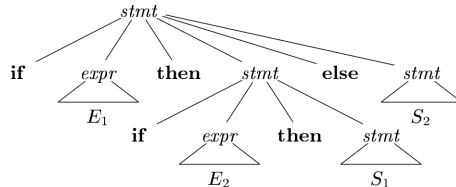
if E_1 then if E_2 then S_1 else S_2

- درخت تجزیه‌ای برای این جمله رسم کنید.

- حال جمله زیر را در نظر بگیرید.

if E_1 then if E_2 then S_1 else S_2

- برای این جمله دو درخت تجزیه به صورت زیر وجود دارد.



- در همه زبان‌های برنامه‌نویسی درخت تجزیه اول را به درخت دوم ترجیح می‌دهیم.



- در واقع قانونی که در همه زبان‌ها وجود دارد این است که *else* به نزدیک‌ترین *if* (یا *then*) قبل از آن تعلق دارد.

- می‌توانیم یک گرامر غیرمبهم به صورت زیر تولید کنیم که معادل گرامر مبهم ذکر شده است.
- توجه کنید که بین `then` و `else` اگر قرار باشد دستور شرطی `if` قرار بگیرد، باید حتماً یک `if-then-else` باشد، در غیراینصورت ابهام به وجود می‌آید.
- در واقع قانونی که برای گرامر غیرمبهم وضع می‌کنیم این است که همیشه بین `then` و `else` یا یک عبارت `if-then-else` قرار می‌گیرد و یا یک دستور غیرشرطی. اما بعد از `else` ممکن است یک عبارت شرطی بدون `else` به کار رفته شود.

- گرامر زیر معادل گرامر شرطی برای دستورات if-then-else است که ابهام در آن رفع شده است.

<i>stmt</i>	→	<i>matched_stmt</i>
		<i>open_stmt</i>
<i>matched_stmt</i>	→	if <i>expr</i> then <i>matched_stmt</i> else <i>matched_stmt</i>
		other
<i>open_stmt</i>	→	if <i>expr</i> then <i>stmt</i>
		if <i>expr</i> then <i>matched_stmt</i> else <i>open_stmt</i>

تجزیه بالا به پایین

- قبل از بررسی مفصل تجزیه‌کنندهٔ بالا به پایین، برای یک گرامر ساده که زیر مجموعه‌ای از گرامر زبان جاوا و سی است، یک تجزیه‌کنندهٔ بالا به پایین¹ می‌سازیم و سپس در مورد روند کلی ساختن تجزیه‌کنندهٔ بالا به پایین صحبت می‌کنیم.
- گرامر زیر را در نظر بگیرید.

$$\begin{array}{lcl} stmt & \rightarrow & \text{expr} ; \\ & | & \text{if (expr) stmt} \\ & | & \text{for (optexpr ; optexpr ; optexpr) stmt} \\ & | & \text{other} \end{array}$$
$$\begin{array}{lcl} optexpr & \rightarrow & \epsilon \\ & | & \text{expr} \end{array}$$

- در اینجا **expr** و **other** را به عنوان دو ترمینال در نظر گرفتیم. در یک گرامر کامل این دو را به عنوان دو متغیر در نظر می‌گیریم و توسط قوانین دیگر تعریف می‌کنیم.

¹ top-down parser

تجزیه بالا به پایین

- تجزیه کننده بالا به پایین یک درخت تجزیه با یک ریشه می‌سازد به طوری که برچسب ریشه درخت متغیر آغازین گرامر ($stmt$) است.
 - تجزیه کننده بالا به پایین به طور خلاصه به طور مکرر عملیات زیر را انجام می‌دهد.
۱. در رأس N با برچسب A ، یکی از قوانین تولید متغیر A را انتخاب می‌کند و فرزندان N را نمادهای (متغیرها و ترمینال‌های) بدنه قانون انتخاب شده قرار می‌دهد.
 ۲. رأس بعدی در درخت تجزیه که با یک متغیر برچسب زده است و چپ‌ترین متغیر در بین همه برگ‌هاست را انتخاب می‌کند و آن برگ را با توجه به رشته ورودی گسترش می‌دهد.
- در هرگام از فرایند تجزیه، تجزیه کننده با توجه به توکن بعدی¹ در رشته ورودی تصمیم می‌گیرد چه قانونی را انتخاب کند.

¹ lookahead

- در تجزیه رشته `other (;expr;expr) for` اولین توکن، واژه `for` است. بنابراین ریشه درخت تجزیه که با `stmt` برچسب زده شده است با قانونی از متغیر `stmt` گسترش می‌یابد که بدنه آن با واژه `for` آغاز شده است.
- در گام بعد در درخت تجزیه باید برگ را تجزیه کنیم که بعد از برگ با برچسب `for` قرار دارد. این برگ (است و در رشته ورودی نیز توکن بعدی نماد) است. در اینجا نماد درخت تجزیه بر نماد رشته ورودی منطبق می‌شود و در درخت تجزیه و رشته ورودی باید به سمت نماد بعدی حرکت کنیم.
- در گام بعد نماد بعدی در درخت تجزیه را انتخاب می‌کنیم که اولین رخداد متغیر `optexpr` است. این روند ادامه می‌یابد تا کل رشته ورودی تجزیه شود.

تجزیه بالا به پایین

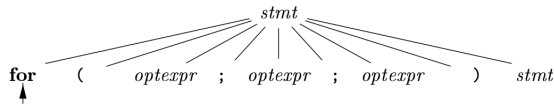
- شکل زیر روند تجزیه یک رشته توسط تجزیه کننده بالا به پایین را نشان می دهد.

PARSE
TREE

stmt
↑

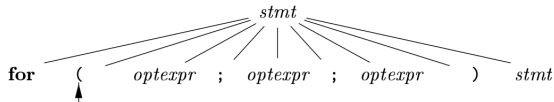
INPUT **for** (; **expr** ; **expr**) **other**
 ↑

PARSE
TREE



INPUT **for** (; **expr** ; **expr**) **other**
 ↑

PARSE
TREE



INPUT **for** (; **expr** ; **expr**) **other**
 ↑

- در حالت کلی در یک تجزیه کننده بالا به پایین ممکن است انتخاب یک قانون با خطا روبرو شود که در این صورت باید با استفاده از پسگرد یا عقبگرد¹ قانون بعدی انتخاب شود.
- معمولا چنین عقبگردهایی پرهزینه است و به دنبال روش‌های تجزیه‌ای هستیم که از چنین عقبگره‌هایی جلوگیری کنند.

¹ backtrack

- تجزیه کاهشی بازگشتی¹ روشی بالا به پایین برای تحلیل نحوی است که در آن مجموعه‌ای از توابع بازگشتی برای پردازش رشته ورودی استفاده می‌شوند. در این تجزیه‌کننده، به ازای هریک از متغیرهای گرامر یک تابع در نظر گرفته می‌شود.
- یکی از انواع ساده تجزیه کاهشی بازگشتی، تجزیه پیش‌بینی‌کننده² است. در تجزیه پیش‌بینی‌کننده از پسگرد جلوگیری می‌شود.

¹ recursive-descent parsing

² predictive parsing

- یک تجزیه‌کننده پیش‌بینی‌کننده برای گرامر قبل در زیر نشان داده شده است.

```
void stmt() {
    switch ( lookahead ) {
        case expr:
            match(expr); match(';'); break;
        case if:
            match(if); match('('); match(expr); match(')'); stmt();
            break;
        case for:
            match(for); match('(');
            optexpr(); match(';'); optexpr(); match(';'); optexpr();
            match(')'); stmt(); break;
        case other:
            match(other); break;
        default:
            report("syntax error");
    }
}

void optexpr() {
    if ( lookahead == expr ) match(expr);
}

void match(terminal t) {
    if ( lookahead == t ) lookahead = nextTerminal;
    else report("syntax error");
}
```

- در این تجزیه‌کننده به ازای هر متغیر گرامر یک تابع تعریف می‌شود. بسته به این که توکن بعدی در رشته ورودی چه مقداری دارد، تجزیه‌کننده، ورودی را با گرامر تطبیق می‌دهد.
- برای تطبیق¹ یک ترمینال در گرامر و یک کلمه از رشته ورودی، تجزیه‌کننده صرفاً بررسی می‌کند که ترمینال گرامر و توکن بعدی در رشته ورودی برابر باشند.
- برای تطبیق یک متغیر در گرامر و یک کلمه از رشته ورودی، تجزیه‌کننده تابع متناظر با متغیر را فراخوانی می‌کند.

¹ match

- تجزیه‌کننده پیش‌بینی‌کننده برای یک گرامر ساده¹ می‌تواند مورد استفاده قرار بگیرد.
- برای تعریف گرامر ساده، تابع $\text{First}(\alpha)$ را به صورت زیر تعریف می‌کنیم، جایی که α دنباله‌ای از ترمینال‌ها و متغیرهاست.
- اگر اولین کلمه در دنباله α ترمینال t باشد، آنگاه $\text{First}(\alpha) = \{t\}$.
- اگر اولین کلمه در دنباله α متغیر A باشد و داشته باشیم $A \rightarrow \beta_1 \mid \beta_2 \mid \dots \mid \beta_n$ آنگاه
$$\text{First}(\alpha) = \text{First}(\beta_1) \cup \text{First}(\beta_2) \cup \dots \cup \text{First}(\beta_n)$$

¹ simple grammar

- حال فرض کنید در گرامر G داشته باشیم $A \rightarrow \alpha$ و $A \rightarrow \beta$. گرامر G ساده است اگر $\text{First}(\alpha) \cap \text{First}(\beta) = \emptyset$.
- از تجزیه‌کننده پیش‌بینی‌کننده تنها زمانی می‌توان استفاده کرد که یک گرامر ساده باشد. در اینصورت در زمان خطی یک رشته را می‌توان تجزیه کرد.

- در تجزیه‌کننده پیش‌بینی‌کننده‌ای که طراحی کردیم، در پیاده‌سازی تابع $\text{optexpr}()$ در صورتی که تطبیق رخ ندهد خطایی صادر نکردیم. با این کار در واقع قانون $\text{optexpr} \rightarrow \epsilon$ را پیاده‌سازی کردیم.
- در حالت کلی اگر قانونی به صورت $\epsilon \mid X_1 \mid X_2 \mid \dots \mid A$ داشته باشیم، در پیاده‌سازی تابع $A()$ هیچ خطایی صادر نمی‌کنیم، زیرا ممکن است رشته ورودی، در بدنه هیچ یک از قوانین A منطبق نشود که در این صورت قانون تهی اعمال می‌شود.

تجزیه پیش‌بینی‌کننده

- برای پیاده‌سازی یک تجزیه‌کننده پیش‌بینی‌کننده برای یک گرامر ساده، به ازای هر یک از متغیرهای گرامر یک تابع تعریف می‌کنیم. با شروع از تابع متعلق به متغیر آغازین تجزیه‌کننده مکرراً به ازای هر متغیر A در گرامر که دارای قوانین $A \rightarrow X_1 \mid X_2 \mid \dots \mid X_n$ است، قانون $A \rightarrow X_i$ را انتخاب می‌کند، اگر توکن بعدی در رشته ورودی در مجموعه $\text{First}(X_i)$ باشد.
- با فرض براینکه X_i دنباله‌ای از ترمینال‌های t و متغیرهای X است، به ازای هر ترمینال t ، توکن بعدی در رشته ورودی باید برابر با ترمینال t باشد و به ازای هر متغیر X ، تابع $X()$ فراخوانی می‌شود.
- اگر رشته ورودی بدون خطا پایان رسید، رشته متعلق به زبان آن گرامر است.

- ممکن است یک تجزیه‌کننده کاهشی بازگشتی¹ در یک حلقه بی‌پایان بیافتد.
- فرض کنید یک قانون بازگشتی چپ² به صورت زیر داشته باشیم :

$$\text{expr} \rightarrow \text{expr} + \text{term}$$

¹ recursive-descent parser

² left-recursive rule

تجزیه پیش‌بینی‌کننده

- در این قانون، متغیر سمت چپ قانون برابر با نماد سمت چپ در بدنه قانون است.
- حال در فرایند تجزیه اگر تابع $\text{expr}()$ فراخوانی شود، این تابع نیز مجدداً تابع $\text{expr}()$ را فراخوانی می‌کند و این فراخوانی بازگشتی خاتمه پیدا نمی‌کند.
- برای استفاده از تجزیه‌کننده کاهشی بازگشتی باید قوانین بازگشتی چپ را حذف کنیم.

حذف بازگشت چپ

- یک گرامر بازگشتی چپ¹ است اگر به ازای متغیر A و صورت جمله‌ای دلخواه α فرایند اشتقاق $A \xRightarrow{+} A\alpha$ وجود داشته باشد.
- تجزیه کننده‌های بالا به پایین نمی‌توانند گرامرهایی که دارای بازگشت چپ هستند را تجزیه کنند، بنابراین بازگشت چپ² باید در گرامر حذف شود.

¹ left recursive

² left recursion

حذف بازگشت چپ

- گرامر زیر بازگشتی چپ است.

$$\begin{aligned} E &\rightarrow E + T \mid T \\ T &\rightarrow T * F \mid F \\ F &\rightarrow (E) \mid \mathbf{id} \end{aligned}$$

- پس از حذف بازگشت چپ در این گرامر، گرامر زیر به دست می‌آید.

$$\begin{aligned} E &\rightarrow T E' \\ E' &\rightarrow + T E' \mid \epsilon \\ T &\rightarrow F T' \\ T' &\rightarrow * F T' \mid \epsilon \\ F &\rightarrow (E) \mid \mathbf{id} \end{aligned}$$

- قوانین تولید $E \rightarrow E + T \mid T$ با قوانین $E \rightarrow T E'$ و $E' \rightarrow + T E' \mid \epsilon$ جایگزین می‌شوند.

حذف بازگشت چپ

- بازگشت چپ بلاواسطه¹ می‌تواند توسط روش زیر حذف شود.
- ابتدا قوانین را به صورت زیر مرتب می‌کنیم.

$$A \rightarrow A\alpha_1 \mid A\alpha_2 \mid \dots \mid A\alpha_m \mid \beta_1 \mid \beta_2 \mid \dots \mid \beta_n$$

به طوری که β_i ها با A آغاز نمی‌شوند.

- سپس قوانین تولید متغیر A را با قوانین زیر جایگزین می‌کنیم.

$$A \rightarrow \beta_1 A' \mid \beta_2 A' \mid \dots \mid \beta_n A'$$

$$A' \rightarrow \alpha_1 A' \mid \alpha_2 A' \mid \dots \mid \alpha_m A' \mid \epsilon$$

- بدین صورت متغیر A همان رشته‌های قبلی را تولید می‌کند با این تفاوت که بازگشت چپ حذف شده است.

¹ immediate left recursion

حذف بازگشت چپ

- یک قانون بازگشتی چپ به صورت $A \rightarrow A\alpha \mid \beta$ در واقع رشته‌هایی به صورت $\beta\alpha^*$ تولید می‌کند. چنین رشته‌هایی را می‌توانیم با گرامر غیر بازگشتی $A \rightarrow \beta R$ و $R \rightarrow \alpha R \mid \epsilon$ نیز تولید کنیم.



- گرامر جایگزین یک گرامر بازگشتی راست¹ است زیرا قانون $R \rightarrow \alpha R$ متغیر R را در سمت راست بدنه قانون دارد. قوانین بازگشتی راست در تجزیه کننده‌های بالا به پایین مشکلی به وجود نمی‌آورند.

¹ right recursive

حذف بازگشت چپ

- روشی که مطرح کردیم بازگشت چپ بلاواسطه را حذف می‌کند اما همه بازگشت‌های چپ را حذف نمی‌کند. برخی مواقع پس از چندگام در فرایند اشتقاق بازگشت چپ به وجود می‌آید.
- برای مثال گرامر زیر را در نظر بگیرید :

$$S \rightarrow Aa \mid b$$

$$A \rightarrow Ac \mid Sd \mid \epsilon$$

- متغیر S بازگشت چپ بلاواسطه ندارد ولی در فرایند اشتقاق خواهیم داشت $S \Rightarrow Aa \Rightarrow Sda$ که یک بازگشت چپ است.
- الگوریتم زیر برای گرامرهایی که در آنها دور وجود ندارد یعنی اشتقاق $A \stackrel{+}{\Rightarrow} A$ اتفاق نمی‌افتد و همچنین در آنها قانون تولید تهی یعنی $A \rightarrow \epsilon$ وجود ندارد، بازگشت چپ را حذف می‌کند.

- الگوریتم حذف بازگشت چپ، گرامر G بدون دور و بدون قانون تولید تهی را دریافت می‌کند و یک گرامر معادل بدون بازگشت چپ تولید می‌کند.

- 1) arrange the nonterminals in some order A_1, A_2, \dots, A_n .
- 2) **for** (each i from 1 to n) {
- 3) **for** (each j from 1 to $i - 1$) {
- 4) replace each production of the form $A_i \rightarrow A_j \gamma$ by the
 productions $A_i \rightarrow \delta_1 \gamma \mid \delta_2 \gamma \mid \dots \mid \delta_k \gamma$, where
 $A_j \rightarrow \delta_1 \mid \delta_2 \mid \dots \mid \delta_k$ are all current A_j -productions
- 5) }
- 6) eliminate the immediate left recursion among the A_i -productions
- 7) }

حذف بازگشت چپ

- این الگوریتم به صورت زیر عمل می‌کند.

- به ازای $i = 1$ دو حالت وجود دارد. یا $A_1 \rightarrow A_1 \alpha$ و یا $A_1 \rightarrow A_m \alpha$ به طوری که $m > 1$. در حالت اول در خط ۶ الگوریتم بازگشت چپ بلاواسطه حذف می‌شود. در حالت دوم قانون به همان صورت باقی خواهد ماند.

- به ازای $i = 2$ سه حالت وجود دارد. یا $A_2 \rightarrow A_2 \alpha$ یا $A_2 \rightarrow A_1 \alpha$ و یا $A_2 \rightarrow A_m \alpha$ به ازای $m > 2$. در حالت اول بازگشت چپ بلاواسطه حذف می‌شود. در حالت دوم قانون به صورت $A_2 \rightarrow A_p \alpha$ به ازای $p > 1$ بازنویسی می‌شود و در صورتی که $p = 2$ باشد، بازگشت چپ بلاواسطه حذف می‌شود. در حالت سوم قانون به همان صورت باقی می‌ماند. پس در پایان در هر صورت خواهیم داشت $A_2 \rightarrow A_m \alpha$ به ازای $m > 2$.

- در حالت کلی پس از اتمام تکرار i ام در الگوریتم، به ازای همه قوانین A_i خواهیم داشت $A_i \rightarrow A_m \alpha$ به طوری که $m > i$.

- در تکرار آخر الگوریتم، بازگشت چپ برای A_n در صورت وجود حذف می‌شود.

حذف بازگشت چپ

- برای استفاده از این الگوریتم ابتدا همه قوانین تولید تهی به صورت $A \rightarrow \epsilon$ را حذف می‌کنیم. اگر قانون $A \rightarrow \epsilon$ وجود داشته باشد، قانون $S \rightarrow AS\alpha$ نیز دارای بازگشت چپ است که این الگوریتم قادر به تشخیص آن نیست.
- توجه کنید که پس از حذف قوانین تولید تهی، تنها در صورتی می‌توانیم $A \xRightarrow{+} A$ داشته باشیم که قوانین یک به صورت $A \rightarrow B$ وجود داشته باشند.
- بنابراین همه قوانین تولید یک به صورت $A \rightarrow B$ را نیز حذف می‌کنیم.
- الگوریتم‌های ساده‌سازی گرامرهای مستقل از متن برای حذف قوانین تولید تهی و یک به در مبحث نظریه زبان‌ها و ماشین‌ها بررسی می‌شوند.

حذف بازگشت چپ

- علت ایجاد اشکال در گرامر با وجود دورها به شرح زیر است.
- فرض کنید داشته باشیم $A_i \rightarrow A_j$ به طوری که $j > i$.
- در بازنویسی قانون $A_j \rightarrow A_i$ خواهیم داشت $A_j \rightarrow A_j$.
- اما برای حذف بازگشت چپ در این قانون با بازگشت چپ مواجه می‌شویم. در واقع اگر داشته باشیم $A \rightarrow A|\beta$ با حذف بازگشت چپ بلاواسطه به دست می‌آوریم $A \rightarrow \beta R, R \rightarrow R|\epsilon$ که همچنان بازگشتی چپ است.

حذف بازگشت چپ

- می‌خواهیم بازگشت چپ را در گرامر زیر حذف کنیم. قانون تولید تهی در اینجا مشکلی در اجرای الگوریتم ایجاد نمی‌کند، اما در حالت کلی قوانین تولید تهی را حذف می‌کنیم.

$$S \rightarrow Aa \mid b$$

$$A \rightarrow Ac \mid Sd \mid \epsilon$$

- متغیرها را به صورت S, A مرتب می‌کنیم. سپس قوانین زیر را تولید می‌کنیم.

$$S \rightarrow Aa \mid b$$

$$A \rightarrow Ac \mid Aad \mid bd \mid \epsilon$$

- با حذف بازگشت‌های چپ بلاواسطه گرامر زیر را به دست می‌آوریم.

$$S \rightarrow Aa \mid b$$

$$A \rightarrow bdA' \mid A'$$

$$A' \rightarrow cA' \mid adA' \mid \epsilon$$

- فاکتورگیری چپ روشی است برای تبدیل کردن یک گرامر به گرامری که برای تجزیه کننده بالا به پایین پیش‌بینی‌کننده مناسب باشد.
 - وقتی برای جایگزین کردن یک متغیر با بدنه قانون در فرایند اشتقاق دو انتخاب داشته باشیم، در مواردی می‌توانیم انتخاب را به تعویق بیندازیم تا وقتی که ورودی بیشتری خوانده شود.
 - برای مثال فرض کنید قوانین تولیدی به صورت زیر داریم.
- $$\begin{array}{lcl} stmt & \rightarrow & \text{if } expr \text{ then } stmt \text{ else } stmt \\ & | & \text{if } expr \text{ then } stmt \end{array}$$
- با خواندن توکن `if` از ورودی نمی‌توانیم تصمیم بگیریم کدام قانون را انتخاب کنیم.

- در حالت کلی اگر دو قانون $A \rightarrow \alpha\beta_1 \mid \alpha\beta_2$ را داشته باشیم و ورودی α باشد، نمی‌توانیم تصمیم بگیریم کدام قانون را انتخاب کنیم، اما می‌توانیم گرامر را به گونه‌ای تغییر دهیم که انتخاب به تعویق بیافتد.
- می‌توانیم این گرامر را به صورت زیر بنویسیم :

$$\begin{aligned} A &\rightarrow \alpha A' \\ A' &\rightarrow \beta_1 \mid \beta_2 \end{aligned}$$

فاکتورگیری چپ

- الگوریتم فاکتورگیری، گرامر G را دریافت می‌کند و گرامری تولید می‌کند که در آن فاکتورگیری چپ اعمال شده باشد.
- برای هر متغیر A ، بلندترین پیشوند α بین دو یا چند انتخاب را پیدا می‌کنیم. اگر $\alpha \neq \epsilon$ آنگاه قوانین $A \rightarrow \alpha\beta_1 \mid \alpha\beta_2 \mid \dots \mid \alpha\beta_n \mid \gamma$ را با قوانین زیر جایگزین می‌کنیم. قوانین γ قوانینی هستند که پیشوند آنها α نیست.

$$A \rightarrow \alpha A' \mid \gamma$$

$$A' \rightarrow \beta_1 \mid \beta_2 \mid \dots \mid \beta_n$$

- این روند را برای همه متغیرها تکرار می‌کنیم.

فاکتورگیری چپ

- گرامر زیر معادل گرامر if-else است. در این گرامر i و t و e نماینده if و then و else هستند.

$$S \rightarrow i E t S \mid i E t S e S \mid a$$

$$E \rightarrow b$$

- می‌توانیم این گرامر را به صورت زیر فاکتورگیری چپ کنیم.

$$S \rightarrow i E t S S' \mid a$$

$$S' \rightarrow e S \mid \epsilon$$

$$E \rightarrow b$$

- توجه کنید که هر دوی این گرامرها مبهم هستند.

ساختارهای غیرمستقل از متن

- برخی از ساختارها در زبان‌های برنامه‌نویسی را نمی‌توان توسط گرامرهای مستقل از متن توصیف کرد.
- برای مثال در بسیاری از زبان‌ها نیاز داریم که متغیر قبل از استفاده تعریف شده باشد.
- این ساختار را می‌توانیم به صورت wcw مدلسازی کنیم جایی که اولین w نماینده تعریف متغیر، c نماینده قسمتی از کد برنامه، و دومین w نماینده استفاده از متغیر باشد.
- می‌توان اثبات کرد که زبان $L = \{wcw \mid w \in (a|b)^*\}$ مستقل از متن نیست.

ساختارهای غیرمستقل از متن

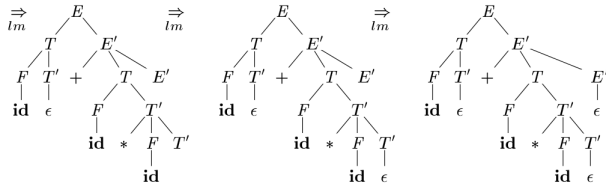
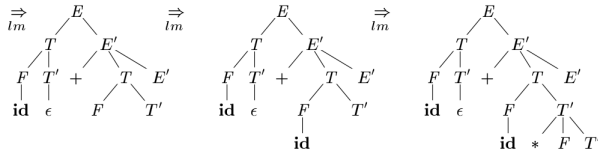
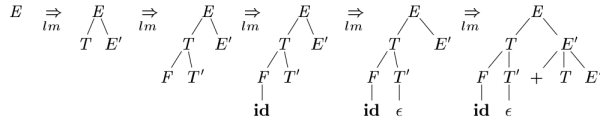
- در نتیجه نیاز به روش‌های دیگر برای تحلیل معنای برنامه‌ها داریم.
- یک مثال دیگر از ساختارهایی از زبان که مستقل از متن نیستند، به شرح زیر است. در زبان‌های برنامه‌نویسی نیاز است که تعداد آرگومان‌های ارسال شده به یک تابع برابر با تعداد پارامترهای تعریف شده در تابع باشد. فرض کنید تعریف دو تابع با n و m ورودی را به صورت a^n و b^m نشان دهیم و دو فراخوانی تابع از این دو تابع را به صورت c^n و d^m .
- این ساختار را با زبان $L = \{a^n b^m c^n d^m \mid n \geq 1, m \geq 1\}$ مدلسازی می‌کنیم. می‌توان اثبات کرد که این زبان مستقل از متن نیست.

تجزیه بالا به پایین

- تجزیه بالا به پایین برای تجزیه یک رشته، درخت تجزیه را با شروع از ریشه می‌سازد.
- برای مثال برای تجزیه رشته $\text{id} + \text{id} * \text{id}$ با استفاده از گرامر زیر، از تجزیه بالا به پایین صفحه بعد استفاده می‌کنیم.

$$\begin{aligned} E &\rightarrow T E' \\ E' &\rightarrow + T E' \mid \epsilon \\ T &\rightarrow F T' \\ T' &\rightarrow * F T' \mid \epsilon \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$

تجزیه بالا به پایین



- ریشه درخت تجزیه متغیر آغازین است. در هرگام، تجزیه کننده باید تصمیم بگیرد از کدام یک از قوانین تولید استفاده کند برای اینکه بتواند رشته مورد نظر را تجزیه کند.
- ابتدا در مورد یک تجزیه کننده به نام تجزیه کننده کاهشی بازگشتی¹ صحبت می‌کنیم که در آن برای پیدا کردن قانون مناسب در فرایند تجزیه از پسگرد² استفاده می‌شود.
- سپس در مورد یک حالت خاص تجزیه کننده کاهشی بازگشتی به نام تجزیه کننده پیش بینی کننده³ صحبت می‌کنیم که در آن به پسگرد نیازی نیست.

¹ recursive-descent parser

² backtrack

³ predictive parser

- تجزیه کننده پیش بینی کننده با بررسی چند نماد بعدی در رشته ورودی تصمیم می گیرد کدام قانون تولید را انتخاب کند و به پسگرد نیازی ندارد.
- گرامرهایی که با بررسی k نماد در ورودی می توانیم برای آنها تجزیه کننده پیش بینی کننده بسازیم، گرامرهای $LL(k)$ نامیده می شوند.

تجزیه کننده کاهشی بازگشتی

- یک تجزیه کننده کاهشی بازگشتی برنامه‌ای است که از مجموعه‌ای از توابع تشکیل شده است به طوری که هر تابع متعلق به یکی از متغیرهای گرامر است. اجرای تجزیه کننده با فراخوانی تابع متعلق به متغیر آغازین شروع می‌شود و در نهایت اگر همه رشته ورودی خوانده شد متوقف می‌شود.

- الگوریتم تجزیه کننده کاهشی بازگشتی در زیر نشان داده شده است.

```
void A() {  
1)      Choose an  $A$ -production,  $A \rightarrow X_1 X_2 \cdots X_k$ ;  
2)      for (  $i = 1$  to  $k$  ) {  
3)          if (  $X_i$  is a nonterminal )  
4)              call procedure  $X_i()$ ;  
5)          else if (  $X_i$  equals the current input symbol  $a$  )  
6)              advance the input to the next symbol;  
7)          else /* an error has occurred */;  
      }  
}
```

تجزیه کننده کاهشی بازگشتی

- یک تجزیه کننده کاهشی بازگشتی با استفاده از یک الگوریتم پسگرد رشته ورودی را تجزیه می‌کند، اما برای تجزیه زبان‌های برنامه‌نویسی معمولاً نیازی به پسگرد نیست.
- برای اینکه در تجزیه بالا به پایین از پسگرد استفاده کنیم، در خط (۱) برنامه قبل باید همه انتخاب‌های موجود برای جایگزینی متغیر A را امتحان کنیم. همچنین در خط (۷) در صورتی که به بن‌بست برخورد کردیم پیام خطا صادر نمی‌کنیم بلکه پسگرد انجام می‌شود.

تجزیه کننده گاهشی بازگشتی

- گرامر زیر را در نظر بگیرید.

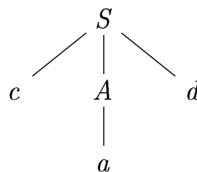
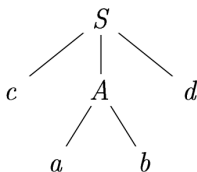
$$S \rightarrow cAd$$

$$A \rightarrow ab \mid a$$

- برای ساختن یک درخت تجزیه از بالا به پایین برای رشته $w = cad$ با ریشه درخت تجزیه یعنی S آغاز می‌کنیم. S تنها یک قانون دارد. بنابراین رأس S را بسط می‌دهیم و فرزندان آن شامل c و A و d را می‌سازیم. اولین برگ یعنی c بر رشته ورودی منطبق می‌شود، پس در رشته ورودی جلو می‌رویم. برگ بعدی A است که یک متغیر با دو قانون است. آن را با استفاده از اولین قانون یعنی $A \rightarrow ab$ بسط می‌دهیم. نماد a در رشته ورودی بر دومین برگ یعنی a منطبق می‌شود پس در رشته ورودی به جلو می‌رویم. اما سومین برگ یعنی b بر نماد بعدی در ورودی یعنی d منطبق نمی‌شود پس باید به عقب برگردیم و یک قانون دیگر از A را انتخاب کنیم. با بازگشت به عقب باید در رشته ورودی هم به عقب برگردیم، پس در هر رأس باید اندیس رشته ورودی ذخیره شود.

تجزیه کننده کاهشی بازگشتی

- روند تجزیه کاهشی بازگشتی و پسگرد به صورت زیر است.



- اگر یک گرامر بازگشت چپ داشته باشد، تجزیه کننده کاهشی بازگشتی وارد حلقه بی پایان می شود.

- برای ساختن تجزیه کننده‌های بالا به پایین و پایین به بالا به دو تابع مهم به نام First و Follow نیاز داریم که در اینجا به آنها اشاره می‌کنیم.
- در هنگام تجزیه بالا به پایین این توابع کمک می‌کنند قانون درست را با استفاده از نماد ورودی بعد انتخاب کنیم.
- در هنگام بازیابی خطا از توکن‌های تولید شده توسط تابع Follow استفاده می‌شود.

- اگر α یک رشته از نمادهای گرامر باشد، آنگاه $\text{First}(\alpha)$ مجموعه‌ای از ترمینال‌هایی است که در ابتدای رشته‌های مشتق شده از α وجود دارند. اگر $\alpha \xRightarrow{*} \epsilon$ آنگاه ϵ نیز در $\text{First}(\alpha)$ است.
- برای مثال در شکل زیر $A \xRightarrow{*} c\gamma$ بنابراین c در $\text{First}(A)$ است.



- تابع First در تجزیه پیش‌بینی کننده استفاده می‌شود. فرض کنید دو قانون $A \rightarrow \alpha|\beta$ را داشته باشیم و $First(\alpha)$ و $First(\beta)$ دو مجموعه مجزا باشند. آنگاه با خواندن ورودی a می‌توانیم قانون مورد نظر برای اعمال را انتخاب کنیم زیرا a می‌تواند حداکثر در یکی از مجموعه‌های $First(\alpha)$ یا $First(\beta)$ باشد.

- به ازای متغیر A تابع $\text{Follow}(A)$ مجموعه ترمینال‌های a است که مستقیماً در سمت راست متغیر A در یک صورت جمله‌ای در یک فرایند اشتقاق قرار می‌گیرند.
- به عبارت دیگر $\text{Follow}(A)$ مجموعه ترمینال‌های a است که برای آنها اشتقاق $\alpha A a \beta$ $S \xRightarrow{*}$ وجود دارد. توجه کنید که بین A و a در فرایند اشتقاق می‌تواند متغیرهایی وجود داشته باشند ولی این متغیرها به تهی تبدیل می‌شوند.
- همچنین اگر A متغیر سمت راست باشد آنگاه $\$$ در $\text{Follow}(A)$ قرار می‌گیرد. نماد $\$$ به معنای پایان رشته است و فرض می‌شود که این نماد در الفبا وجود ندارد.

توابع First و Follow

- برای محاسبه $First(X)$ برای نماد X قوانین زیر را اعمال می‌کنیم تا جایی که هیچ ترمینالی (یا رشته ϵ) نتواند به مجموعه $First(X)$ اضافه شود.

۱. اگر X یک ترمینال است آنگاه $First(X) = \{X\}$.

۲. اگر X یک متغیر است و $X \rightarrow Y_1 Y_2 \dots Y_k$ یک قانون تولید است به ازای $k \geq 1$ آنگاه $First(X)$ در $First(Y_i)$ قرار می‌گیرد اگر به ازای یک i دلخواه، ترمینال a در $First(Y_i)$ باشد و ϵ در همه مجموعه‌های $First(Y_1), \dots, First(Y_{i-1})$ باشد. در اینصورت خواهیم داشت $\epsilon \xRightarrow{*} Y_1 \dots Y_{i-1}$. اگر به ازای همه $1 \leq j \leq k$ ، رشته ϵ در $First(Y_j)$ باشد آنگاه ϵ را به $First(X)$ اضافه می‌کنیم. برای مثال، هر نمادی در $First(Y_1)$ است در $First(X)$ نیز وجود دارد. اگر Y_1 رشته تهی ϵ را مشتق نمی‌کند آنگاه نماد دیگری به $First(X)$ اضافه نمی‌کنیم، اما اگر $\epsilon \xRightarrow{*} Y_1$ آنگاه $First(Y_2)$ را نیز به $First(X)$ می‌افزاییم و به همین ترتیب الی آخر.

۳. اگر $X \rightarrow \epsilon$ یک قانون تولید باشد، آنگاه ϵ را به $First(X)$ می‌افزاییم.

- می‌توانیم $\text{First}(X_1X_2 \cdots X_n)$ را به صورت زیر محاسبه کنیم. همه نمادها غیر از ϵ از مجموعه $\text{First}(X_1)$ را به $\text{First}(X_1X_2 \cdots X_n)$ می‌افزاییم. اگر ϵ در مجموعه $\text{First}(X_1)$ باشد، آنگاه همه نمادهای غیر ϵ از $\text{First}(X_2)$ را به $\text{First}(X_1X_2 \cdots X_n)$ می‌افزاییم. اگر ϵ در $\text{First}(X_1)$ و $\text{First}(X_2)$ باشد آنگاه این همه نمادهای غیر از ϵ از $\text{First}(X_3)$ را به $\text{First}(X_1X_2 \cdots X_n)$ می‌افزاییم. این روند را ادامه می‌دهیم. در نهایت اگر ϵ به ازای همه i ها در $\text{First}(X_i)$ باشد، آنگاه ϵ را به $\text{First}(X_1X_2 \cdots X_n)$ اضافه می‌کنیم.

- برای محاسبه $\text{Follow}(A)$ به ازای همه متغیرهای A قوانین زیر را اعمال می‌کنیم تا وقتی که هیچ نمادی نتواند به مجموعه $\text{Follow}(A)$ اضافه شود.

۱. اگر S متغیر آغازین باشد، نماد $\$$ را در $\text{Follow}(S)$ اضافه می‌کنیم.

۲. اگر قانون $A \rightarrow \alpha B \beta$ وجود داشته باشد، آنگاه همه نمادهای مجموعه $\text{First}(\beta)$ به جز رشته تهی ϵ را به $\text{Follow}(B)$ می‌افزاییم.

۳. اگر قانون $A \rightarrow \alpha B \beta$ یا قانون $A \rightarrow \alpha B \beta$ وجود داشته باشد جایی که $\text{First}(\beta)$ حاوی ϵ باشد، آنگاه هر نمادی در $\text{Follow}(A)$ در $\text{Follow}(B)$ نیز قرار می‌گیرد.

توابع First و Follow

- برای مثال گرامر زیر را در نظر بگیرید.

$$S \rightarrow Am \mid An \mid kA \mid Bp$$

$$A \rightarrow qrB \mid s$$

$$B \rightarrow t$$

- با محاسبه توابع First و Follow خواهیم داشت:

$$\text{First}(S) = \{ k, q, s, t \}$$

$$\text{First}(A) = \{ q, s \}$$

$$\text{First}(B) = \{ t \}$$

$$\text{Follow}(S) = \{ \$ \}$$

$$\text{Follow}(A) = \{ \$, m, n \}$$

$$\text{Follow}(B) = \{ \$, m, n, p \}$$

- گرامر زیر را در نظر بگیرید.

$$\begin{aligned} E &\rightarrow T E' \\ E' &\rightarrow + T E' \mid \epsilon \\ T &\rightarrow F T' \\ T' &\rightarrow * F T' \mid \epsilon \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$

- با محاسبات توابع First و Follow به دست می آوریم :

$$\text{First}(F) = \text{First}(T) = \text{First}(E) = \{ (, \text{id} \}$$

$$\text{First}(E') = \{ +, \epsilon \}$$

$$\text{First}(T') = \{ *, \epsilon \}$$

$$\text{Follow}(E) = \text{Follow}(E') = \{), \$ \}$$

$$\text{Follow}(T) = \text{Follow}(T') = \{ +,), \$ \}$$

$$\text{Follow}(F) = \{ +, *,), \$ \}$$

گرامرهای $LL(1)$

- تجزیه کننده‌های پیش‌بینی کننده یعنی تجزیه کننده‌های کاهشی بازگشتی که به پسگرد نیازی ندارند می‌توانند برای دسته‌ای از گرامرهای مستقل از متن به نام گرامرهای $LL(1)$ استفاده شوند.
- تجزیه کننده‌هایی که برای گرامرهای $LL(1)$ به کار می‌روند تجزیه کننده‌های $LL(1)$ نامیده می‌شوند. اولین L بدین معناست که خواندن ورودی از چپ به راست¹ انجام می‌شود و دومین L بدین معناست که تجزیه کننده اشتقاق چپ² تولید می‌کند و عدد ۱ بدین معناست که تجزیه کننده تنها یک نماد جلوتر³ را در هر گام برای تصمیم‌گیری برای تجزیه بررسی می‌کند.
- دسته گرامرهای $LL(1)$ برای توصیف زبان‌های برنامه‌نویسی به اندازه کافی توانمند است. البته در توصیف یک گرامر $LL(1)$ ملاحظات را باید در نظر گرفت. برای مثال گرامر نباید بازگشت چپ داشته باشد یا مبهم باشد.

¹ Left to right

² Leftmost derivation

³ one input symbol of lookahead

- گرامر مستقل از متن G یک گرامر $LL(1)$ است اگر و تنها اگر هنگامی که دو قانون مجزای $A \rightarrow \alpha | \beta$ وجود داشته باشند، شرط‌های زیر برقرار باشد.

۱. هیچ ترمینال a وجود ندارد به طوری که هر دوی α و β رشته‌ای مشتق کنند، که هر دو با a آغاز شود.

۲. حداکثر یکی از صورت‌های جمله‌ای α و β می‌توانند رشته تهی تولید کنند.

۳. اگر $\epsilon \Rightarrow^* \beta$ آنگاه α هیچ رشته‌ای تولید نمی‌کند که با یک ترمینال در $\text{Follow}(A)$ آغاز شود. به طور

مشابه اگر $\epsilon \Rightarrow^* \alpha$ آنگاه β هیچ رشته‌ای تولید نمی‌کند که با یک ترمینال در $\text{Follow}(A)$ آغاز شود.

فرض کنید این شرایط برقرار نباشد و داشته باشیم $a \Rightarrow^* \beta a \Rightarrow^* Aa \Rightarrow^* S$ و همچنین

$a\gamma a \Rightarrow^* \alpha a \Rightarrow^* Aa \Rightarrow^* S$. در این صورت نمی‌توانیم تصمیم بگیریم با مشاهده توکن a در ورودی در فرایند اشتقاق برای متغیر A کدامیک از قوانین $A \rightarrow \alpha$ یا $A \rightarrow \beta$ را انتخاب کنیم.

- شرط‌های اول و دوم معادل یکدیگرند. این دو شرط بدین معنی هستند که $First(\alpha)$ و $First(\beta)$ دو مجموعه مجزا هستند.
- شرط سوم بدین معنی است که اگر ϵ در $First(\beta)$ وجود داشت، آنگاه $First(\alpha)$ و $Follow(A)$ دو مجموعه مجزا هستند و به طور مشابه اگر ϵ در $First(\alpha)$ وجود داشت، $First(\beta)$ و $Follow(A)$ دو مجموعه مجزا هستند.

- تجزیه کننده‌های پیش‌بینی کننده می‌توانند برای گرامرهای LL(۱) استفاده شوند زیرا انتخاب درست قانونی که می‌تواند در هر گام برای تجزیه به کار رود تنها با بررسی نماد بعدی در ورودی امکان پذیر است.
- برای مثال در گرامر زیر تنها با خواندن یکی از نمادهای if یا while یا { می‌توانیم تصمیم بگیریم کدام قانون را انتخاب کنیم.

```
stmt → if (expr) stmt else stmt  
      | while (expr) stmt  
      | { stmt-list }
```

الگوریتم تجزیه کننده پیش‌بینی کننده

- الگوریتم بعدی اطلاعاتی در مورد مجموعه‌های First و Follow در یک جدول تجزیه پیش‌بینی کننده جمع‌آوری می‌کند. جدول تجزیه $M[A, a]$ یک آرایه دو بعدی است جایی که A یک متغیر و a یک ترمینال یا نماد $\$$ است.
- الگوریتم بر پایه ایده زیر است : قانون $A \rightarrow \alpha$ انتخاب می‌شود اگر نماد بعدی a در $\text{First}(\alpha)$ باشد. در صورتی که $\alpha = \epsilon$ یا $\alpha \xRightarrow{*} \epsilon$ آنگاه $A \rightarrow \alpha$ را انتخاب می‌کنیم اگر نماد ورودی در $\text{Follow}(A)$ باشد یا اگر به $\$$ در رشته ورودی رسیده‌ایم و $\$$ در $\text{Follow}(A)$ باشد.

الگوریتم تجزیه کننده پیش‌بینی کننده

- الگوریتم ساخت جدول تجزیه کننده پیش‌بینی کننده به صورت زیر است. این الگوریتم گرامر G را دریافت و جدول تجزیه M را تولید می‌کند.
- برای هر یک از قوانین $A \rightarrow \alpha$ از گرامر به صورت زیر عمل می‌کنیم.
 ۱. به ازای هریک از ترمینال‌های a در $\text{First}(\alpha)$ قانون $A \rightarrow \alpha$ را به $M[A, a]$ اضافه می‌کنیم.
 ۲. اگر ϵ در $\text{First}(\alpha)$ باشد، آنگاه به ازای هریک از ترمینال‌های a در $\text{Follow}(A)$ قانون $A \rightarrow \alpha$ را به $M[A, a]$ اضافه می‌کنیم. اگر ϵ در $\text{First}(\alpha)$ باشد و $\$$ در $\text{Follow}(A)$ باشد آنگاه $A \rightarrow \alpha$ را به $M[A, \$]$ اضافه می‌کنیم.
- اگر پس از عملیات بالا هیچ قانونی در $M[A, a]$ قرار نگرفت، آنگاه در $M[A, a]$ مقدار خطا (error) قرار می‌دهیم. برای سادگی، خطاها را با خانه‌های خالی در جدول نمایش می‌دهیم.

الگوریتم تجزیه کننده پیش‌بینی کننده

- برای گرامر زیر جدول تجزیه زیر تولید می‌شود.

$$\begin{aligned} E &\rightarrow T E' \\ E' &\rightarrow + T E' \mid \epsilon \\ T &\rightarrow F T' \\ T' &\rightarrow * F T' \mid \epsilon \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$

NON - TERMINAL	INPUT SYMBOL					
	id	+	*	()	\$
E	$E \rightarrow T E'$			$E \rightarrow T E'$		
E'		$E' \rightarrow + T E'$			$E' \rightarrow \epsilon$	$E' \rightarrow \epsilon$
T	$T \rightarrow F T'$			$T \rightarrow F T'$		
T'		$T' \rightarrow \epsilon$	$T' \rightarrow * F T'$		$T' \rightarrow \epsilon$	$T' \rightarrow \epsilon$
F	$F \rightarrow \text{id}$			$F \rightarrow (E)$		

الگوریتم تجزیه کننده پیش‌بینی کننده

- قانون $E \rightarrow TE'$ را در نظر بگیرید. از آنجایی که $\text{First}(TE') = \{ (, \text{id} \}$ این قانون به $M[E, (]$ و $M[E, \text{id}]$ افزوده شده است.
- قانون $E' \rightarrow +TE'$ به $M[E', +]$ افزوده شده است زیرا $\text{First}(+TE') = \{ + \}$.
- از آنجایی که $\text{Follow}(E') = \{), \$ \}$ قانون $E' \rightarrow \epsilon$ به $M[E',)]$ و $M[E', \$]$ افزوده شده است.

NON - TERMINAL	INPUT SYMBOL					
	id	+	*	()	\$
E	$E \rightarrow TE'$			$E \rightarrow TE'$		
E'		$E' \rightarrow +TE'$			$E' \rightarrow \epsilon$	$E' \rightarrow \epsilon$
T	$T \rightarrow FT'$			$T \rightarrow FT'$		
T'		$T' \rightarrow \epsilon$	$T' \rightarrow *FT'$		$T' \rightarrow \epsilon$	$T' \rightarrow \epsilon$
F	$F \rightarrow \text{id}$			$F \rightarrow (E)$		

الگوریتم تجزیه کننده پیش‌بینی کننده

- الگوریتمی که شرح داده شد می‌تواند بر روی هر گرامر G اعمال شود و یک جدول تجزیه M بسازد. برای هر گرامر $LL(1)$ یک جدول تجزیه وجود دارد که هر خانه آن حاوی یک قانون تولید یا خطا است.
- برای برخی از گرامرها، جدول M ممکن است خانه‌ای داشته باشد که در آن بیش از یک قانون وجود دارد. اگر یک گرامر بازگشت چپ داشته باشد یا مبهم باشد، آنگاه M حداقل یک خانه با بیش از یک قانون دارد. با حذف بازگشت چپ و فاکتورگیری چپ در برخی موارد می‌توان یک گرامر را به یک گرامر $LL(1)$ تبدیل کرد. اما برای برخی از گرامرها معادل $LL(1)$ وجود ندارد.

الگوریتم تجزیه کننده پیش‌بینی کننده

- گرامر زیر و جدول تجزیه آن را در نظر بگیرید.

$$S \rightarrow iEtSS' \mid a$$

$$S' \rightarrow eS \mid \epsilon$$

$$E \rightarrow b$$

NON - TERMINAL	INPUT SYMBOL					
	a	b	e	i	t	$\$$
S	$S \rightarrow a$			$S \rightarrow iEtSS'$		
S'			$S' \rightarrow \epsilon$ $S' \rightarrow eS$			$S' \rightarrow \epsilon$
E		$E \rightarrow b$				

- در سلول $M[S', e]$ دو قانون $S' \rightarrow eS$ و $S' \rightarrow \epsilon$ قرار گرفته است.

- دلیل این امر این است که گرامر مبهم است.

تجزیه کننده پیش‌بینی کننده غیر بازگشتی

- یک تجزیه کننده پیش‌بینی کننده غیر بازگشتی¹ با نگهداری یک پشته به صورت صریح به جای استفاده از پشته فراخوانی ساخته می‌شود.
- این تجزیه کننده اشتقاق چپ را شبیه‌سازی می‌کند.
- اگر w رشته ورودی باشد که بر گرامر تطبیق داده شده باشد، آنگاه پشته یک دنباله از نمادهای α از گرامر را نگهداری می‌کند به طوری که $S \xRightarrow{*}_{lm} w\alpha$.
- تجزیه کننده تشکیل شده است از یک بافر ورودی، یک پشته حاوی دنباله‌ای از نمادهای گرامر، یک جدول تجزیه که توسط الگوریتم قبل ساخته شده است و یک خروجی.

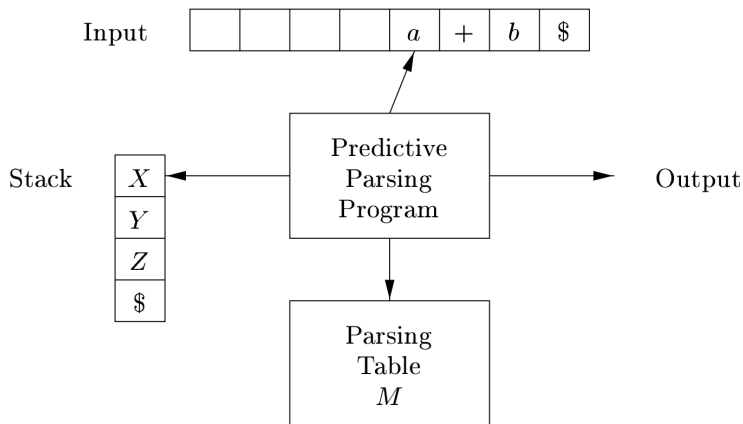
¹ nonrecursive predictive parser

تجزیه کننده پیش بینی کننده غیر بازگشتی

- بافر ورودی رشته ورودی را در بر می گیرد که با نماد \$ ختم شده است. همچنین نماد \$ انتهای رشته را نشان می دهد.
- تجزیه کننده نماد X را از روی رشته بر می دارد و نماد a را از ورودی می خواند. اگر X یک متغیر باشد، تجزیه کننده قانون تولیدی را که در $M[X, a]$ ذخیره شده انتخاب می کند. در غیر این صورت نماد X و نماد ورودی a باید تطبیق داده شوند.

تجزیه کننده پیش بینی کننده غیر بازگشتی

- شمای این تجزیه کننده در زیر نشان داده شده است.



تجزیه کننده پیش بینی کننده غیربازگشتی

- الگوریتم رشته w و جدول M برای گرامر G را دریافت می کند. اگر w در $L(G)$ باشد یک اشتقاق چپ برای w تولید می کند در غیراینصورت پیام خطا صادر می کند.
- در ابتدا تجزیه کننده در پیکربندی $^1 \$w$ قرار دارد و نماد آغازین S در پشته بر روی نماد انتهای پشته یعنی $\$$ قرار گرفته می شود.

¹ configuration

تجزیه کننده پیش‌بینی کننده غیر بازگشتی

- الگوریتم زیر عملیات تجزیه پیش‌بینی کننده را نشان می‌دهد.

```
let  $a$  be the first symbol of  $w$ ;  
let  $X$  be the top stack symbol;  
while (  $X \neq \$$  ) { /* stack is not empty */  
    if (  $X = a$  ) pop the stack and let  $a$  be the next symbol of  $w$ ;  
    else if (  $X$  is a terminal )  $error()$ ;  
    else if (  $M[X, a]$  is an error entry )  $error()$ ;  
    else if (  $M[X, a] = X \rightarrow Y_1 Y_2 \cdots Y_k$  ) {  
        output the production  $X \rightarrow Y_1 Y_2 \cdots Y_k$ ;  
        pop the stack;  
        push  $Y_k, Y_{k-1}, \dots, Y_1$  onto the stack, with  $Y_1$  on top;  
    }  
    let  $X$  be the top stack symbol;  
}
```

تجزیه کننده پیش بینی کننده غیر بازگشتی

- گرامر زیر را در نظر بگیرید.

$$\begin{array}{lcl} E & \rightarrow & T E' \\ E' & \rightarrow & + T E' \mid \epsilon \\ T & \rightarrow & F T' \\ T' & \rightarrow & * F T' \mid \epsilon \\ F & \rightarrow & (E) \mid \mathbf{id} \end{array}$$

تجزیه کننده پیش بینی کننده غیر بازگشتی

- جدول تجزیه این گرامر را قبلا به صورت زیر محاسبه کردیم.

NON - TERMINAL	INPUT SYMBOL					
	id	+	*	()	\$
E	$E \rightarrow TE'$			$E \rightarrow TE'$		
E'		$E' \rightarrow +TE'$			$E' \rightarrow \epsilon$	$E' \rightarrow \epsilon$
T	$T \rightarrow FT'$			$T \rightarrow FT'$		
T'		$T' \rightarrow \epsilon$	$T' \rightarrow *FT'$		$T' \rightarrow \epsilon$	$T' \rightarrow \epsilon$
F	$F \rightarrow \text{id}$			$F \rightarrow (E)$		

تجزیه کننده پیش بینی کننده غیر بازگشتی

- با دریافت ورودی $id + id * id$ تجزیه کننده پیش بینی کننده غیر بازگشتی یک فرایند اشتقاق چپ به صورت زیر تولید می کند.

$$E \xRightarrow{lm} TE' \xRightarrow{lm} FT'E' \xRightarrow{lm} idT'E' \xRightarrow{lm} idE' \xRightarrow{lm} id + TE' \xRightarrow{lm} \dots$$

تجزیه کننده پیش بینی کننده غیر بازگشتی

- این فرایند اشتقاق به صورت زیر تولید می شود.

MATCHED	STACK	INPUT	ACTION
	$E\$$	$\text{id} + \text{id} * \text{id}\$$	
	$TE'\$$	$\text{id} + \text{id} * \text{id}\$$	output $E \rightarrow TE'$
	$FT'E'\$$	$\text{id} + \text{id} * \text{id}\$$	output $T \rightarrow FT'$
	$\text{id } T'E'\$$	$\text{id} + \text{id} * \text{id}\$$	output $F \rightarrow \text{id}$
id	$T'E'\$$	$+ \text{id} * \text{id}\$$	match id
id	$E'\$$	$+ \text{id} * \text{id}\$$	output $T' \rightarrow \epsilon$
id	$+ TE'\$$	$+ \text{id} * \text{id}\$$	output $E' \rightarrow + TE'$
$\text{id} +$	$TE'\$$	$\text{id} * \text{id}\$$	match $+$
$\text{id} +$	$FT'E'\$$	$\text{id} * \text{id}\$$	output $T \rightarrow FT'$
$\text{id} +$	$\text{id } T'E'\$$	$\text{id} * \text{id}\$$	output $F \rightarrow \text{id}$
$\text{id} + \text{id}$	$T'E'\$$	$* \text{id}\$$	match id
$\text{id} + \text{id}$	$* FT'E'\$$	$* \text{id}\$$	output $T' \rightarrow * FT'$
$\text{id} + \text{id} *$	$FT'E'\$$	$\text{id}\$$	match $*$
$\text{id} + \text{id} *$	$\text{id } T'E'\$$	$\text{id}\$$	output $F \rightarrow \text{id}$
$\text{id} + \text{id} * \text{id}$	$T'E'\$$	$\$$	match id
$\text{id} + \text{id} * \text{id}$	$E'\$$	$\$$	output $T' \rightarrow \epsilon$
$\text{id} + \text{id} * \text{id}$	$\$$	$\$$	output $E' \rightarrow \epsilon$

تجزیه کننده پیش‌بینی کننده غیربازگشتی

– یک صورت جمله‌ای در فرایند اشتقاق متناظر است با ورودی تطبیق داده شده (در ستون Matched) که به دنبال آن محتوای پشته قرار داده شده است.

بازیابی خطا در تجزیه کننده پیش‌بینی کننده

- یک خطا در تجزیه پیش‌بینی کننده رخ می‌دهد وقتی که یک ترمینال بر روی پشته بر روی نماد ورودی منطبق نشود و یا وقتی که با خواندن متغیر A از پشته و نماد a از رشته ورودی، $M[A, a]$ یک خطا باشد یا به عبارت دیگر خانه $M[A, a]$ در جدول تجزیه خالی باشد.

بازیابی خطا با توکن همگام‌کننده

- بازیابی خطا با توکن همگام‌کننده¹ بر این پایه است که در هنگام رخداد خطا از نمادهای ورودی چشم‌پوشی شود تا جایی که یک توکن همگام‌کننده² پیدا شود.
- مجموعه توکن‌های همگام‌کننده باید به نحوی انتخاب شود که تجزیه‌کننده بتواند به سرعت خطا را بازیابی کند.

¹ panic mode error recovery

² synchronizing token

بازیابی خطا با توکن همگام‌کننده

- از قوانین زیر می‌توان در بازیابی خطا با توکن همگام‌کننده استفاده کرد.

۱. همه نمادها در $\text{Follow}(A)$ را در مجموعه همگام‌کننده متغیر A قرار می‌دهیم. اگر از همه توکن‌ها چشم‌پوشی کنیم تا یکی از اعضای $\text{Follow}(A)$ مشاهده شود و A از پشته خارج شود، به احتمال زیاد تجزیه می‌تواند ادامه پیدا کند.

- برای مثال فرض کنید قوانین $S \rightarrow E; S|E$ و $E \rightarrow \text{id } E' \text{ و } E' \rightarrow + \text{id } E' | \epsilon$ در یک گرامر وجود داشته باشد. برای تجزیه جمله $\text{id}(\text{id} + \text{id}; \text{id} + \text{id}; \text{id} + \text{id}; S\$)$ وقتی در پشته نمادهای $E'; S\$$ قرار داشته باشند و به عبارت $(\text{id} + \text{id}; \text{id} + \text{id}; \text{id} + \text{id}; \text{id})$ برخورد کنیم، از کاراکترها چشم‌پوشی می‌شود تا اینکه به یک کاراکتر نقطه ویرگول برخورد کنیم، زیرا $\text{Follow}(E') = \{;\}$. سپس E' از پشته خارج می‌شود و نقطه ویرگول در پشته تطبیق داده می‌شود و تجزیه با عبارت $\text{id} + \text{id};$ ادامه پیدا می‌کند.

بازیابی خطا با توکن همگام‌کننده

- تنها اعضای $\text{Follow}(A)$ برای مجموعه همگام‌کننده A کافی نیستند. برای مثال اگر دستورات با نقطه ویرگول خاتمه پیدا کنند، آنگاه کلمه‌های کلیدی که در ابتدای دستورات بعدی هستند در مجموعه Follow قرار نمی‌گیرند. بنابراین اگر یک نقطه ویرگول جا افتاده باشد، از کلمات کلیدی دستورات بعدی چشم‌پوشی می‌شود. معمولاً در زبان‌های برنامه‌نویسی یک ساختار سلسله مراتبی وجود دارد. برای مثال عبارات در دستورات استفاده می‌شوند و دستورات در بلوک‌ها و الی آخر. می‌توانیم نمادهایی را که متغیرهای سلسله‌مراتب بالاتر با آنها آغاز می‌شوند، به مجموعه همگام‌کننده از متغیرهای سلسله مراتب پایین‌تر اضافه کنیم. برای مثال، می‌توانیم کلمات کلیدی را که دستورات با آنها شروع می‌شوند در مجموعه‌های همگام‌کننده متغیرهایی قرار دهیم که عبارات را تولید می‌کنند.
- برای مثال فرض کنید قوانین $S \rightarrow E; S | E; \text{int id}$ و $E \rightarrow \text{id } E' | \text{int id}$ و $E' \rightarrow + \text{id } E' | \epsilon$ در یک گرامر وجود داشته باشد. از آنجایی که E می‌تواند با کلمه int آغاز شود، آن را به مجموعه همگام‌کننده E' می‌افزاییم. اگر در پشته نمادهای $E'; S$ قرار داشته باشند و به عبارت $(\text{id} + \text{id } \text{int id};$ برخورد کنیم، از کاراکترها چشم‌پوشی می‌شود تا اینکه به یک کاراکتر نقطه ویرگول یا توکن int برخورد کنیم، زیرا $\text{synch}(E') = \{;, \text{int}\}$.

بازیابی خطا با توکن همگام‌کننده

۲. در صورتی که در ورودی با خطا روبرو شدیم و در پشته متغیر A قرار داشته باشد، می‌توانیم از توکن‌ها چشم‌پوشی کنیم تا به یک کاراکتر در $\text{First}(A)$ برخورد کنیم. آنگاه می‌توانیم تجزیه را با توجه به متغیر A ادامه دهیم، اگر یک نماد در $\text{First}(A)$ در ورودی ظاهر شود.

- برای مثال فرض کنید قوانین $S \rightarrow E; S | E$ و $E \rightarrow \text{id } E' \mid \epsilon$ و $E' \rightarrow + \text{id } E' \mid \epsilon$ در یک گرامر وجود داشته باشد. در تجزیه $((+ \text{id}; S)$ وقتی در پشته نمادهای $(E'; S)$ قرار داشته باشند و به عبارت $((+ \text{id}; S)$ برخورد کنیم، از کاراکترهای $(($ چشم‌پوشی می‌شود و یک پیام خطا صادر می‌شود تا اینکه به یک کاراکتر $+$ برخورد کنیم، زیرا $\text{First}(E') = \{+\}$.

بازیابی خطا با توکن همگام‌کننده

۳. اگر یک ترمینال بر روی پشته باشد که نتواند تطبیق داده شود، می‌توان ترمینال را از روی پشته برداشته، خطایی صادر کرد مبنی بر اینکه ترمینال توسط کامپایلر اضافه شده است و عملیات تجزیه را ادامه داد.

- برای مثال فرض کنید قوانین $S \rightarrow E; S|E$ و $E \rightarrow id\ E'$ و $E' \rightarrow +\ id\ E'|\epsilon$ در یک گرامر وجود داشته باشد. در تجزیه $id\ id$ وقتی در پشته نمادهای $+idE; S$ قرار داشته باشند و به عبارت id برخورد کنیم، توکن $+$ از پشته برداشته می‌شود و یک پیام خطا صادر می‌شود، مبنی بر اینکه توکن $+$ در ورودی فراموش شده است، و تجزیه ادامه پیدا می‌کند.

بازیابی خطا با توکن همگام‌کننده

- گرامر زیر و جدول تجزیه متناظر با آن را در نظر بگیرید.

$$\begin{aligned} E &\rightarrow T E' \\ E' &\rightarrow + T E' \mid \epsilon \\ T &\rightarrow F T' \\ T' &\rightarrow * F T' \mid \epsilon \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$

NON - TERMINAL	INPUT SYMBOL					
	id	+	*	()	\$
E	$E \rightarrow TE'$			$E \rightarrow TE'$		
E'		$E' \rightarrow +TE'$			$E' \rightarrow \epsilon$	$E' \rightarrow \epsilon$
T	$T \rightarrow FT'$			$T \rightarrow FT'$		
T'		$T' \rightarrow \epsilon$	$T' \rightarrow *FT'$		$T' \rightarrow \epsilon$	$T' \rightarrow \epsilon$
F	$F \rightarrow \text{id}$			$F \rightarrow (E)$		

بازیابی خطا با توکن همگام‌کننده

- در جدول زیر واژه synch در خانه $M[A, a]$ قرار گرفته است، اگر $a \in \text{Follow}(A)$ باشد.

NON - TERMINAL	INPUT SYMBOL					
	id	+	*	()	\$
E	$E \rightarrow TE'$			$E \rightarrow TE'$	synch	synch
E'		$E \rightarrow +TE'$			$E \rightarrow \epsilon$	$E \rightarrow \epsilon$
T	$T \rightarrow FT'$	synch		$T \rightarrow FT'$	synch	synch
T'		$T' \rightarrow \epsilon$	$T' \rightarrow *FT'$		$T' \rightarrow \epsilon$	$T' \rightarrow \epsilon$
F	$F \rightarrow \text{id}$	synch	synch	$F \rightarrow (E)$	synch	synch

بازیابی خطا با توکن همگام‌کننده

- از این جدول به صورت زیر استفاده می‌شود.
- (قانون ۱) اگر تجزیه کننده به کلمه `synch` برخورد کرد، متغیر از روی پشته برداشته می‌شود و پیام خطا صادر می‌شود تا تجزیه بتواند ادامه پیدا کند.
- (قانون ۲) اگر تجزیه کننده به سلول $M[A, a]$ رسید که خالی بود آنگاه پیام خطا صادر شده، از a چشم‌پوشی می‌شود.
- (قانون ۳) اگر یک توکن از روی پشته بر نماد ورودی تطبیق داده نشود، آنگاه توکن از پشته برداشته می‌شود و پیام خطا صادر می‌شود مبنی بر اینکه توکن مورد نظر اضافه شده است.

بازیابی خطا با توکن همگام‌کننده

- با خواندن ورودی $+id$ * id * تجزیه کننده به صورت زیر عمل می‌کند.

STACK	INPUT	REMARK
$E \$$	$* id * + id \$$	error, skip $*$
$E \$$	$id * + id \$$	id is in $FIRST(E)$
$TE' \$$	$id * + id \$$	
$FT' E' \$$	$id * + id \$$	
$id T' E' \$$	$id * + id \$$	
$T' E' \$$	$* + id \$$	
$* FT' E' \$$	$* + id \$$	
$FT' E' \$$	$+ id \$$	error, $M[F, +] = \text{synch}$
$T' E' \$$	$+ id \$$	F has been popped
$E' \$$	$+ id \$$	
$+ TE' \$$	$+ id \$$	
$TE' \$$	$id \$$	
$FT' E' \$$	$id \$$	
$id T' E' \$$	$id \$$	
$T' E' \$$	$\$$	
$E' \$$	$\$$	
$\$$	$\$$	

بازیابی خطا با توکن همگام‌کننده

– معمولاً یک کامپایلر خوب پیام‌های خطایی صادر می‌کند که اطلاعات مفیدی به دست برنامه‌نویس می‌دهد.

بازیابی خطا با جایگزینی توکن‌ها

- بازیابی خطا با جایگزینی توکن‌ها¹ بدین صورت پیاده‌سازی می‌شود که به جای سلول‌های خالی در جدول تجزیه، توابعی قرار می‌گیرند که بازیابی خطا را انجام می‌دهند. این توابع می‌توانند نمادهایی را تغییر دهند یا اضافه کنند و یا حذف کنند و پیام خطای مناسب صادر کنند. همچنین این توابع می‌توانند از پشته نمادهایی را خارج کنند یا نمادهایی را جایگزین کنند و یا نمادهایی را به پشته اضافه کنند. باید اطمینان حاصل شود که این توابع ایجاد حلقه بی‌پایان نمی‌کنند.

¹ phrase-level error recovery

– یک تجزیه کننده پایین به بالا¹ درخت تجزیه برای یک ورودی را از برگ‌ها (پایین) به سمت ریشه (بالا) می‌سازد.

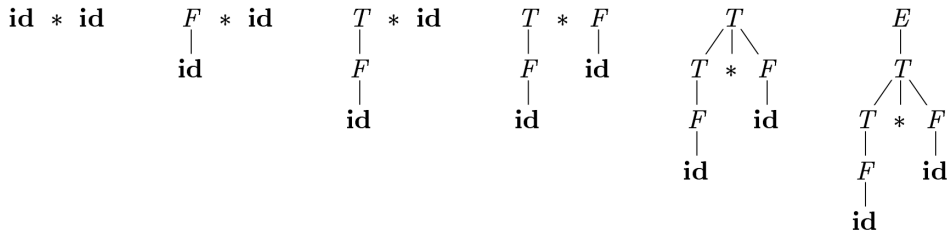
¹ bottom-up parser

تجزیه پایین به بالا

- فرض کنید می‌خواهیم رشته $id * id$ را با استفاده از یک تجزیه کننده پایین به بالا برای گرامر زیر تجزیه کنیم.

$$\begin{aligned} E &\rightarrow E + T \mid T \\ T &\rightarrow T * F \mid F \\ F &\rightarrow (E) \mid id \end{aligned}$$

- فرایند اشتقاق و ساخت درخت تجزیه از پایین به بالا برای این رشته به صورت زیر خواهد بود.



تجزیه پایین به بالا

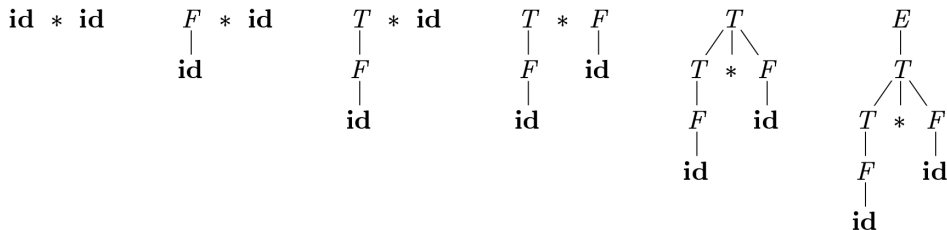
- در این قسمت یک روش کلی برای تجزیه پایین به بالا به نام تجزیه انتقال کاهش¹ معرفی می‌کنیم.
- یکی از دسته‌های مهم گرامرها که برای آنها تجزیه کننده انتقال کاهش می‌تواند ساخته شود، دسته گرامرهای LR نامیده می‌شود.

¹ shift-reduce parsing

- یک تجزیه کننده پایین به بالا با دریافت یک رشته ورودی آن را به متغیر آغازین کاهش می دهد. در هر گام کاهش¹ ، یک زیر رشته از ورودی بر بدنه یک قانون تولید تطبیق پیدا می کند و به متغیر آن قانون تولید کاهش پیدا می کند. یک تجزیه کننده پایین به بالا تعیین می کند کدام قسمت از رشته ورودی توسط کدام یک از قوانین تولید کاهش پیدا کند.

¹ reduction

- در شکل زیر $id * id$ به $F * id$ کاهش پیدا می‌کند و سپس به ترتیب به $T * id$ ، $T * F$ ، T و در نهایت رشته ورودی به E کاهش پیدا می‌کند.



- در گام اول برای کاهش از قانون $F \rightarrow id$ استفاده می‌شود. در برخی از گام‌ها چند انتخاب برای کاهش وجود دارد که تجزیه کنند باید تصمیم بگیرند از کدام قانون و کدام زیر رشته برای کاهش استفاده کنند.

- فرایند کاهش معکوس فرایند اشتقاق است. در فرایند اشتقاق یک متغیر در یک صورت جمله‌ای با بدنه یک قانون از آن متغیر جایگزین می‌شود. اما در فرایند کاهش یک زیررشته از صورت جمله‌ای بر بدنه یک قانون منطبق و با متغیر متعلق به آن قانون جایگزین می‌شود. بنابراین تجزیه کننده پایین به بالا یک اشتقاق به صورت معکوس می‌سازد.

- فرایند اشتقاق راست زیر را در نظر بگیرید.

$$E \Rightarrow T \Rightarrow T * F \Rightarrow T * id \Rightarrow F * id \Rightarrow id * id$$

- در شکل زیر این فرایند اشتقاق از آخر به اول انجام می‌شود.

id * id

F
|
id

T
|
 F
|
id

T
|
 F
|
id

T
/ | \
 T * F
| |
 F **id**
|
id

E
|
 T
/ | \
 T * F
| |
 F **id**
|
id

- در تجزیه پایین به بالا ورودی از چپ به راست خوانده می‌شود و یک اشتقاق راست¹ به صورت معکوس تولید می‌شود.
- یک هندل²، زیر رشته‌ای است که بر بدنه یکی از قوانین تولید تطبیق داده می‌شود. کاهش یک هندل به معنای جایگزین کردن آن با متغیر قانون تولید انتخاب شده است. کاهش هندل یک گام در فرایند اشتقاق راست معکوس است.

¹ rightmost derivation

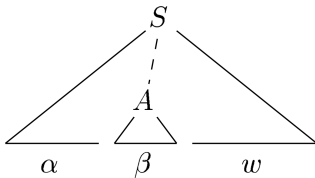
² handle

- برای مثال در جدول زیر هندل در هر گام از فرایند کاهش مشخص شده است. برای خوانایی بیشتر و تمیز دادن توکن‌های id ، از اندیس استفاده شده است.

RIGHT SENTENTIAL FORM	HANDLE	REDUCING PRODUCTION
$\text{id}_1 * \text{id}_2$	id_1	$F \rightarrow \text{id}$
$F * \text{id}_2$	F	$T \rightarrow F$
$T * \text{id}_2$	id_2	$F \rightarrow \text{id}$
$T * F$	$T * F$	$T \rightarrow T * F$
T	T	$E \rightarrow T$

- در این فرایند کاهش در گام سوم گرچه T در $T * \text{id}_2$ می‌تواند به عنوان هندل انتخاب شود چرا که قانون $E \rightarrow T$ وجود دارد، اما T به عنوان هندل انتخاب نمی‌شود چرا که در این صورت متغیر آغازین E به دست نمی‌آید و رشته تجزیه نمی‌شود. همچنین در گام چهارم می‌توانستیم F را به T کاهش دهیم، اما در آن صورت رشته تجزیه نمی‌شد.

- اگر داشته باشیم $S \xrightarrow[rm]{*} \alpha A w \xRightarrow[rm]{} \alpha \beta w$ آنگاه قانون تولید $A \rightarrow \beta$ یک هندل برای صورت جمله‌ای $\alpha \beta w$ در فرایند کاهش است.



- توجه کنید که در تعریف بالا w تنها از ترمینال‌ها تشکیل شده است. برای سادگی به جای $A \rightarrow \beta$ می‌گوییم β یک هندل برای $\alpha \beta w$ است.
- اگر یک گرامر مبهم باشد، چند هندل در فرایند کاهش وجود خواهد داشت.

- فرایند اشتقاق راست معکوس با کاهش هندل به دست می‌آید. در این فرایند با یک رشته w از ترمینال‌ها آغاز می‌کنیم. اگر w یک رشته از گرامر باشد، آنگاه $w = \gamma_n$ جایی که γ_n برابر است با n امین صورت جمله‌ای در فرایند اشتقاق راست.

$$S = \gamma_0 \xRightarrow{rm} \gamma_1 \xRightarrow{rm} \cdots \xRightarrow{rm} \gamma_{n-1} \xRightarrow{rm} \gamma_n = w$$

- برای ساختن این اشتقاق به صورت معکوس، هندل β_n در γ_n پیدا می‌شود و با A_n با استفاده از قانون $A_n \rightarrow \beta_n$ جایگزین می‌شود تا صورت جمله γ_{n-1} به دست بیاید.
- الگوریتم تجزیه پایین به بالا روشی برای یافتن هندل توصیف می‌کند. این فرایند ادامه پیدا می‌کند تا در نهایت متغیر S به دست بیاید. در این صورت رشته تجزیه شده است و متعلق به گرامر است.

- دقت کنید در فرایند اشتقاق راست ابتدا قسمت راست صورت‌های جمله‌ای تجزیه می‌شود، یعنی در یک فرایند اشتقاق داریم $\alpha A w \Rightarrow \alpha \beta w$ جایی که $A \rightarrow \beta$.
- بنابراین اگر بخواهیم صورت جمله‌ای $\alpha \beta w$ را کاهش بدهیم، به $\alpha A w$ می‌رسیم.
- در هیچ فرایند اشتقاق راست، اشتقاق $\alpha A \gamma B w \Rightarrow \alpha \beta \gamma B w$ ممکن نیست.
- بنابراین هیچ فرایند کاهشی وجود ندارد که یک صورت جمله‌ای به صورت $\alpha \beta \gamma B w$ را به $\alpha A \gamma B w$ کاهش دهد.
- به عبارت دیگر در فرایند کاهش αx کاهش در انتهای صورت جمله‌ای α انجام می‌شود.

تجزیه انتقال کاهش

- تجزیه انتقال کاهش نوعی تجزیه پایین به بالا است که در آن یک پشته نمادهای گرامر را نگهداری می‌کند و در بافر ورودی باقیمانده رشته ورودی برای تجزیه مشخص شده است.
- هندل همیشه بر روی پشته قرار گرفته است.
- در انتهای پشته و همچنین در انتهای رشته ورودی علامت \$ را قرار می‌دهیم.
- در ابتدای وضعیت پشته و رشته ورودی به صورت زیر است.

STACK
\$

INPUT
 w \$

تجزیه انتقال کاهش

- تجزیه کننده ورودی را از چپ به راست می خواند و تعداد صفر یا بیشتر نماد از ورودی را در پشته قرار می دهد تا وقتی که یک هندل β بر روی پشته برای کاهش یافت شود. سپس β از پشته حذف می شود و با متغیر قانونی که کاهش با استفاده از آن انجام می شود جایگزین می شود.
- این فرایند ادامه پیدا می کند تا اینکه یا تجزیه کننده با خطا روبرو شود و یا در پشته متغیر آغازین قرار بگیرد و ورودی به پایان برسد. در این صورت وضعیت پشته و رشته ورودی به صورت زیر است و رشته ورودی به درستی تجزیه شده است.

STACK
\$ S

INPUT
\$

تجزیه انتقال کاهش

- شکل زیر گام‌های یک تجزیه کننده انتقال کاهش را برای تجزیه رشته $\text{id} * \text{id}$ نشان می‌دهد.

STACK	INPUT	ACTION
\$	$\text{id}_1 * \text{id}_2$ \$	shift
\$ id_1	$* \text{id}_2$ \$	reduce by $F \rightarrow \text{id}$
\$ F	$* \text{id}_2$ \$	reduce by $T \rightarrow F$
\$ T	$* \text{id}_2$ \$	shift
\$ $T *$	id_2 \$	shift
\$ $T * \text{id}_2$	\$	reduce by $F \rightarrow \text{id}$
\$ $T * F$	\$	reduce by $T \rightarrow T * F$
\$ T	\$	reduce by $E \rightarrow T$
\$ E	\$	accept

تجزیه انتقال کاهش

- در فرایند تجزیه انتقال کاهش چهار عملیات می تواند توسط تجزیه کننده اجرا شود.

۱. انتقال¹ : یک نماد از ورودی به روی پشته انتقال پیدا می کند.

۲. کاهش² : یک هندل که نماد سمت راست آن در بالای پشته است و نماد سمت چپ آن در پشته قرار دارد مشخص می شود و با استفاده از یک قانون گرامر کاهش پیدا می کند. هندل از پشته حذف و متغیر مربوط تولیدکننده هندل بر روی پشته اضافه می شود. هندل همیشه بالای پشته قرار می گیرد نه در وسط آن.

۳. پذیرش³ : عملیات تجزیه به اتمام رسیده و رشته پذیرفته شده است.

۴. خطا⁴ : یک خطا نحوی تشخیص داده شده و یک تابع بازیابی خطا فراخوانی می شود.

¹ shift

² reduce

³ accept

⁴ error

تعارض در تجزیه انتقال کاهش

- برای برخی از گرامرهای مستقل از متن تجزیه انتقال کاهش نمی‌تواند استفاده شود. در چنین گرامرهایی تجزیه کننده انتقال کاهش به یک پیکربندی می‌رسد که در آن تجزیه کننده نمی‌تواند تصمیم بگیرد عملیات انتقال انجام دهد و یا عملیات کاهش. به این شرایط تعارض انتقال کاهش¹ گفته می‌شود. همچنین ممکن است تجزیه کننده نتواند تصمیم بگیرد از بین چند کاهش کدام یک را اعمال کند. به این شرایط تعارض کاهش کاهش² گفته می‌شود.

¹ shift/reduce conflict

² reduce/reduce conflict

تعارض در تجزیه انتقال‌کاهش

- اگر یک تجزیه‌کننده انتقال‌کاهش با آگاهی از k نماد بعدی در ورودی و آگاهی از محتوای پشته بتواند تصمیم بگیرد انتقال را اعمال کند و یا کاهش و بتواند تصمیم درستی درمورد انتخاب عملیات کاهش بگیرد گرامر مورد تجزیه یک گرامر $LR(k)$ نامیده می‌شود.
- گرامری که تجزیه‌کننده انتقال‌کاهش برای تجزیه جملات با آن تنها نیاز به آگاهی از یک نماد بعدی در ورودی داشته باشد، یک گرامر $LR(1)$ نامیده می‌شود.
- حرف L بدین معنی است که ورودی از چپ به راست¹ خوانده می‌شود و حرف R بدین معنی است که یک اشتقاق راست به صورت معکوس² ایجاد می‌شود و k بدین معنی است که آگاهی از k نماد بعدی از ورودی برای تجزیه جمله کافی است.

¹ left-to-right

² rightmost derivation in reverse

تعارض در تجزیه انتقال کاهش

- یک گرامر مبهم هیچ‌گاه نمی‌تواند LR باشد.
- گرامر زیر را در نظر بگیرید.

$$\begin{array}{lcl} stmt & \rightarrow & \text{if } expr \text{ then } stmt \\ & | & \text{if } expr \text{ then } stmt \text{ else } stmt \\ & | & \text{other} \end{array}$$

- اگر پیکربندی زیر را در هنگام تجزیه داشته باشیم، نمی‌توانیم تصمیم بگیریم آیا $\text{if } expr \text{ then } stmt$ همدل است یا خیر.

STACK	INPUT
$\dots \text{if } expr \text{ then } stmt$	$\text{else } \dots \$$

- در اینجا یک تعارض انتقال کاهش به وجود می‌آید.
- تجزیه انتقال کاهش می‌تواند با کمی تغییرات برای گرامرهای مبهم نیز استفاده شود.

تعارض در تجزیه انتقال کاهش

- در برخی مواقع تجزیه کننده نمی تواند تصمیم بگیرد از بین چند هندل کدام یک را انتخاب کند و کدام قانون تولید را در فرایند کاهش اعمال کند.
- فرض کنید گرامری به صورت زیر داریم که در آن فراخوانی تابع و تعریف آرایه شبیه به یکدیگر تعریف می شوند و هردو از نماد پرانتز استفاده می کنند

- | | | | |
|-----|-----------------------|---|--|
| (1) | <i>stmt</i> | → | id (<i>parameter_list</i>) |
| (2) | <i>stmt</i> | → | <i>expr</i> := <i>expr</i> |
| (3) | <i>parameter_list</i> | → | <i>parameter_list</i> , <i>parameter</i> |
| (4) | <i>parameter_list</i> | → | <i>parameter</i> |
| (5) | <i>parameter</i> | → | id |
| (6) | <i>expr</i> | → | id (<i>expr_list</i>) |
| (7) | <i>expr</i> | → | id |
| (8) | <i>expr_list</i> | → | <i>expr_list</i> , <i>expr</i> |
| (9) | <i>expr_list</i> | → | <i>expr</i> |

تعارض در تجزیه انتقال کاهش

- حال یک عبارت به صورت $p(i, j)$ در ورودی پس از تحلیل لغوی به صورت $id(id, id)$ تبدیل می‌شود و به تجزیه‌کننده تحویل داده می‌شود. پس از انتقال سه توکن بر روی پشته، تجزیه‌کننده انتقال کاهش در وضعیت زیر قرار می‌گیرد.

STACK	INPUT
$\dots id (id$	$, id) \dots$

- در اینجا دو قانون ۵ برای کاهش id به پارامتر تابع و قانون ۷ برای کاهش id به نام پارامتر آرایه می‌توانند استفاده شوند و یک تعارض کاهش کاهش به وجود می‌آید.
- یک راه حل این است که به جای id در قانون تولید ۱ از توکن $procid$ استفاده شود. این راه حل تحلیل‌گر لغوی را پیچیده می‌کند، زیرا تحلیل‌گر باید با استفاده از جدول علائم تشخیص دهد یک شناسه نام تابع است یا نام آرایه.
- یک راه حل دیگر تغییر ساختار نحوی برنامه و تغییر زبان برنامه‌نویسی است.

تجزیه LR ساده

- بسیاری از تجزیه‌کننده‌های پایین به بالا بر مبنای تجزیه $LR(k)$ هستند.
- حرف L بدین معناست که ورودی از چپ به راست¹ خوانده می‌شود و حرف R بدین معناست که تجزیه با استفاده از یک فرایند اشتقاق راست معکوس² انجام می‌شود و k به معنای تعداد نمادهای بعدی است که برای تصمیم‌گیری در فرایند تجزیه استفاده می‌شود.
- برای تجزیه زبان‌های برنامه‌نویسی معمولاً از تجزیه‌کننده $LR(1)$ استفاده می‌شود.
- وقتی از تجزیه‌کننده LR صحبت می‌کنیم منظور تجزیه‌کننده $LR(1)$ است.
- ابتدا در مورد یک تجزیه‌کننده LR ساده³ یا SLR صحبت می‌کنیم و سپس با روش‌های پیچیده‌تر تجزیه از جمله تجزیه‌کننده LR استاندارد⁴ یا CLR و تجزیه‌کننده $LALR$ آشنا می‌شویم.

¹ left-to-right

² rightmost derivation in reverse

³ Simple LR (SLR)

⁴ Canonical LR (CLR)

- تجزیه‌کننده‌های LR شبیه به تجزیه‌کننده‌ها LL از یک جدول تجزیه استفاده می‌کنند.
- گرامرهایی که می‌توان برای آنها یک تجزیه‌کننده LR طراحی کرد، گرامرهای LR نامیده می‌شوند.

تجزیه LR ساده

- تجزیه LR به چند دلیل پرکاربرد است :

۱. تجزیه‌کننده‌های LR می‌توانند همهٔ ساختارهای زبان‌های برنامه‌نویسی را که برای آنها یک گرامر مستقل از متن وجود دارد تجزیه کنند. گرامرهای مستقل از متنی وجود دارند که LR نیستند اما این گرامرها در زبان‌های برنامه‌نویسی استفاده نمی‌شوند.
 ۲. تجزیه‌کننده LR روشی است غیربازگشتی برای پیاده‌سازی تجزیه انتقال‌کاهش و در عین حال به اندازه بقیه روش‌های تجزیه کاراست. گرامر مورد تجزیه نیاز به حذف بازگشت چپ ندارد و در برخی مواقع می‌تواند مبهم نیز باشد.
 ۳. روش‌های کارایی برای بازیابی خطا در تجزیه‌کننده LR وجود دارد.
 ۴. دسته گرامرهای LR ابر مجموعه دسته گرامرهای LL است، پس تجزیه‌کننده LR تعداد گرامرهای بیشتری را پوشش می‌دهد.
- تنها عیب تجزیه‌کننده LR این است که ساختن آن بسیار پیچیده است.

- چگونه یک تجزیه‌کننده انتقال‌کاهش تشخیص می‌دهد چه زمانی انتقال و چه زمانی کاهش انجام دهد؟
- برای مثال اگر محتوای پشته T باشد و نماد بعدی $*$ باشد، تجزیه‌کننده چگونه تشخیص می‌دهد T هندل نیست و باید به جای کاهش انتقال انجام دهد؟

STACK	INPUT	ACTION
\$	id ₁ * id ₂ \$	shift
\$ id ₁	* id ₂ \$	reduce by $F \rightarrow \mathbf{id}$
\$ F	* id ₂ \$	reduce by $T \rightarrow F$
\$ T	* id ₂ \$	shift
\$ T *	id ₂ \$	shift
\$ T * id ₂	\$	reduce by $F \rightarrow \mathbf{id}$
\$ T * F	\$	reduce by $T \rightarrow T * F$
\$ T	\$	reduce by $E \rightarrow T$
\$ E	\$	accept

- یک تجزیه‌کننده LR تصمیم انتقال یا کاهش را با نگهداری تعدادی حالت انجام می‌دهد.
- هر یک از این حالت‌ها، مجموعه‌ای از آیتم‌ها¹ را در بر می‌گیرند.
- یک آیتم $LR(0)$ از گرامر G یک قانون تولید گرامر G است که تعدادی نقطه در بین نمادهای بدنه آن افزوده شده‌اند.
- بنابراین قانون $A \rightarrow XYZ$ چهار آیتم به صورت زیر دارد :

$$A \rightarrow \cdot XYZ$$

$$A \rightarrow X \cdot YZ$$

$$A \rightarrow XY \cdot Z$$

$$A \rightarrow XYZ \cdot$$
- قانون تولید $A \rightarrow \epsilon$ تنها یک آیتم به صورت $A \rightarrow \cdot$ دارد.

¹ item

- به طور شهودی، یک آیتم نشان می‌دهد چه مقداری از یک قانون تولید در هر لحظه خوانده شده است.
- برای مثال آیتم $XYZ \rightarrow \cdot A$ نشان دهنده این است که می‌توانیم رشته ورودی را از XYZ مشتق کنیم.
- آیتم $A \rightarrow X \cdot YZ$ نشان دهنده این است که ورودی خوانده شده از X مشتق شده و ممکن است بتوانیم ادامه رشته را از YZ مشتق کنیم.
- آیتم $A \rightarrow XYZ \cdot$ نشان دهنده این است که رشته خوانده شده از XYZ مشتق شده و می‌توانیم XYZ را به A کاهش دهیم.

- یک گروه از مجموعه آیتم‌های $LR(\circ)$ ¹ یک گروه $LR(\circ)$ استاندارد² نامیده می‌شود که از آن برای ساختن یک ماشین متناهی قطعی برای تصمیم‌گیری در فرایند ترجمه استفاده می‌شود.
- این ماشین را ماشین $LR(\circ)$ ³ می‌نامیم.

¹ collection of sets of $LR(\circ)$ items

² canonical $LR(\circ)$ collection

³ $LR(\circ)$ automaton

- هر حالت از ماشین $LR(0)$ نشان دهنده مجموعه‌ای از آیتم‌ها در گروه $LR(0)$ استاندارد¹ است.
- برای ساختن گروه $LR(0)$ استاندارد برای یک گرامر، یک گرامر افزوده شده و دو تابع Closure و Goto را تعریف می‌کنیم.
- اگر G یک گرامر با نماد آغازین S باشد، آنگاه G' یک گرامر افزوده شده² برای G است که نماد آغازین S' و قانون $S' \rightarrow S$ در آن افزوده شده است.

¹ canonical $LR(0)$ collection

² augmented grammar

- اگر I مجموعه‌ای از آیتم‌ها بر روی گرامر G باشد، آنگاه $\text{Closure}(I)$ مجموعه‌ای از آیتم‌ها از I است که با دو قانون زیر ساخته شده‌اند :

۱. هر آیتم در I در $\text{Closure}(I)$ نیز اضافه می‌شود.

۲. اگر $A \rightarrow \alpha \cdot B\beta$ در $\text{Closure}(I)$ باشد و $B \rightarrow \gamma$ یک قانون تولید باشد، آنگاه آیتم $\alpha \cdot \gamma$ در $\text{Closure}(I)$ در صورتی که در $\text{Closure}(I)$ وجود نداشته باشد، به آن اضافه می‌شود. این کار تکرار می‌شود تا جایی که هیچ آیتم دیگری را نتوان به $\text{Closure}(I)$ اضافه کرد.

- به طور شهودی $A \rightarrow \alpha \cdot B\beta$ در $\text{Closure}(I)$ نشان دهنده این است که در یکی از گام‌ها در فرایند تجزیه، زیررشته‌ای که باید تجزیه شود از $B\beta$ مشتق می‌شود. زیر رشته‌ای که از $B\beta$ مشتق یک پیشوند دارد که از B مشتق می‌شود بنابراین یکی از قوانین B باید اعمال شود. بنابراین آیتم‌ها برای همه قوانین متعلق به B را اضافه می‌کنیم. پس اگر $B \rightarrow \gamma$ یک قانون تولید باشد، $B \rightarrow \gamma$ در $\text{Closure}(I)$ قرار می‌گیرد.

- مثال : گرامر زیر را در نظر بگیرید.

$$\begin{aligned} E' &\rightarrow E \\ E &\rightarrow E + T \mid T \\ T &\rightarrow T * F \mid F \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$

- اگر I مجموعه‌ای از یک آیتم $\{[E' \rightarrow \cdot E]\}$ باشد، آنگاه $\text{Closure}(I)$ مجموعه آیتم‌های I_0 را شامل می‌شود.

$$\begin{array}{l} I_0 \\ E' \rightarrow \cdot E \\ E \rightarrow \cdot E + T \\ E \rightarrow \cdot T \\ T \rightarrow \cdot T * F \\ T \rightarrow \cdot F \\ F \rightarrow \cdot (E) \\ F \rightarrow \cdot \text{id} \end{array}$$

- در ابتدا آیت $E \rightarrow \cdot E$ براساس قانون اول در $\text{Closure}(I)$ قرار می‌گیرد.
- از آنجایی که E سمت راست نقطه قرار گرفته است، همه قوانین E را با یک نقطه در سمت چپ بدنه قانون می‌افزاییم: $E \rightarrow \cdot E + T$ و $E \rightarrow \cdot T$.
- در آیت افزوده شده، T پس از نقطه قرار گرفته است، پس قوانین $T \rightarrow \cdot T * F$ و $T \rightarrow \cdot F$ را می‌افزاییم.
- در پایان چون F پس از نقطه قرار گرفته است قوانین متعلق به F را با یک نقطه در سمت چپ بدنه قانون می‌افزاییم پس دو آیت $F \rightarrow \cdot (E)$ و $F \rightarrow \cdot id$ اضافه می‌شوند.

- تابع Closure را می‌توان براساس الگوریتم زیر تولید کرد.

```

SetOfItems CLOSURE( $I$ ) {
     $J = I$ ;
    repeat
        for ( each item  $A \rightarrow \alpha \cdot B \beta$  in  $J$  )
            for ( each production  $B \rightarrow \gamma$  of  $G$  )
                if (  $B \rightarrow \cdot \gamma$  is not in  $J$  )
                    add  $B \rightarrow \cdot \gamma$  to  $J$ ;
    until no more items are added to  $J$  on one round;
    return  $J$ ;
}

```

- مجموعه‌های آیت‌ها را به دو دسته تقسیم می‌کنیم.

۱. آیت‌های هسته ¹: آیت شروع $S \rightarrow S'$ و همه آیت‌هایی که نقطه در سمت چپ بدنه آنها نیست.

۲. آیت‌های غیرهسته ²: همه آیت‌هایی که نقطه در سمت چپ بدنه آنهاست به جز $S \rightarrow S'$.

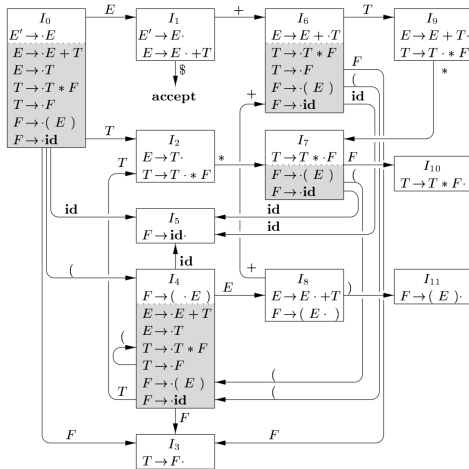
- هر مجموعه از آیت‌ها تشکیل شده است از Closure بر روی مجموعه آیت‌های هسته.

- بنابراین برای صرفه‌جویی در حافظه می‌توانیم آیت‌های غیرهسته را دور بریزیم زیرا این آیت‌ها مجدداً می‌توانند از آیت‌های هسته محاسبه شوند.

¹ kernel items

² nonkernel items

- در شکل زیر آیتم‌های غیرهسته با رنگ خاکستری نشان داده شده‌اند.



- یک تابع مهم دیگر تابع $Goto(I, X)$ است. ورودی I مجموعه‌ای از آیتم‌هاست و X یک نماد از گرامر است.
- اگر در آیتم I داشته باشیم $[A \rightarrow \alpha \cdot X\beta]$ آنگاه تابع $Goto(I, X)$ برابر است با Closure بر روی مجموعه همه آیتم‌های $[A \rightarrow \alpha X \cdot \beta]$.
- به طور شهودی، تابع $Goto$ برای تعریف گذارها در ماشین $LR(0)$ برای یک گرامر استفاده می‌شود. حالت‌های ماشین مجموعه‌ای از آیتم‌هاست و $Goto(I, X)$ گذار از حالت I با ورودی X است.

- اگر I مجموعه‌ای از دو آیت $\{[E' \rightarrow E \cdot], [E \rightarrow E \cdot + T]\}$ باشد، آنگاه $Goto(I, +)$ شامل آیت‌های زیر است.

$$E \rightarrow E + \cdot T$$

$$T \rightarrow \cdot T * F$$

$$T \rightarrow \cdot F$$

$$F \rightarrow \cdot (E)$$

$$F \rightarrow \cdot \text{id}$$

- برای محاسبه $Goto(I, +)$ همه آیت‌هایی را که در آنها $+$ پس از نقطه قرار می‌گیرند در نظر می‌گیریم. آیت $E \rightarrow E \cdot + T$ چنین آیتی است. نقطه را به بعد از $+$ منتقل می‌کنیم و Closure آن را محاسبه می‌کنیم.

- الگوریتم زیر یک گروه استاندارد¹ از مجموعه‌های آیتم‌های LR(°) را برای گرامر افزوده شده G' می‌سازد.

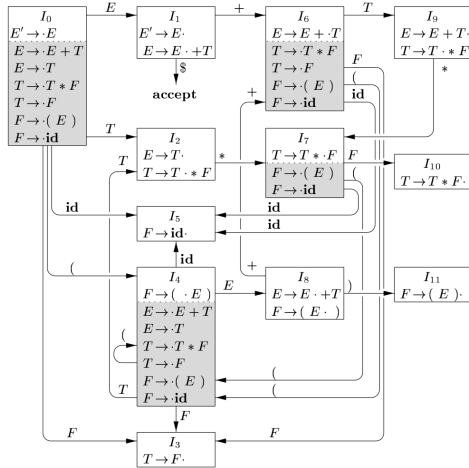
```

void items( $G'$ ) {
     $C = \{\text{CLOSURE}(\{[S' \rightarrow \cdot S]\})\};$ 
    repeat
        for ( each set of items  $I$  in  $C$  )
            for ( each grammar symbol  $X$  )
                if ( GOTO( $I, X$ ) is not empty and not in  $C$  )
                    add GOTO( $I, X$ ) to  $C$ ;
    until no new sets of items are added to  $C$  on a round;
}

```

¹ canonical collection

- گروه استاندارد از مجموعه‌های آیت‌های LR(0) برای گرامر ذکر شده در شکل زیر نشان داده شده است. تابع Closure آیت‌های خاکستری و تابع Goto گذارهای بین حالت‌ها را محاسبه می‌کنند.

$$\begin{aligned} E' &\rightarrow E \\ E &\rightarrow E + T \mid T \\ T &\rightarrow T * F \mid F \\ F &\rightarrow (E) \mid \text{id} \end{aligned}$$


استفاده از ماشین $LR(0)$

- تجزیه LR ساده یا SLR ¹ از ماشین $LR(0)$ استفاده می‌کند.
- حالت‌های ماشین $LR(0)$ مجموعه‌هایی از آیتم‌های گروه‌های $LR(0)$ استاندارد² است و گذارها با تابع Goto محاسبه شده‌اند.

¹ simple LR

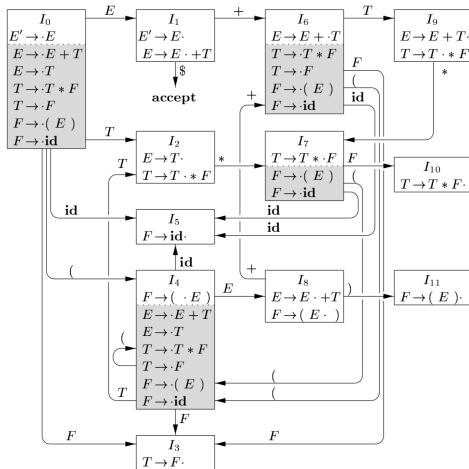
² sets of items from canonical $LR(0)$ collection

استفاده از ماشین $LR(0)$

- حالت آغازین ماشین $LR(0)$ در واقع $Closure([S' \rightarrow \cdot S])$ است، جایی که S' نماد آغازین گرامر افزوده شده است. همهٔ حالت‌ها حالت پذیرش هستند.
- وقتی می‌گوییم حالت z منظور مجموعه آیت‌های I_z است.
- حال باید ببینیم ماشین $LR(0)$ چگونه کمک می‌کند برای انتقال کاهش تصمیم بگیریم.
- فرض کنید رشته γ تشکیل شده از نمادهای گرامر، ماشین $LR(0)$ را از حالت 0 به حالت z می‌برد. حال بر روی نماد ورودی a انتقال انجام می‌دهیم اگر حالت z یک گذار با a دارد. در غیر این صورت یک کاهش انجام می‌دهیم که در این صورت آیت‌های حالت z کمک می‌کنند تصمیم بگیریم از کدام قانون برای کاهش استفاده کنیم.
- الگوریتم تجزیه LR از یک پشته برای نگهداری حالت‌ها استفاده می‌کند.

استفاده از ماشین LR(\circ)

- رشته ورودی $\text{id} * \text{id}$ و ماشین LR(\circ) زیر را در نظر بگیرید.

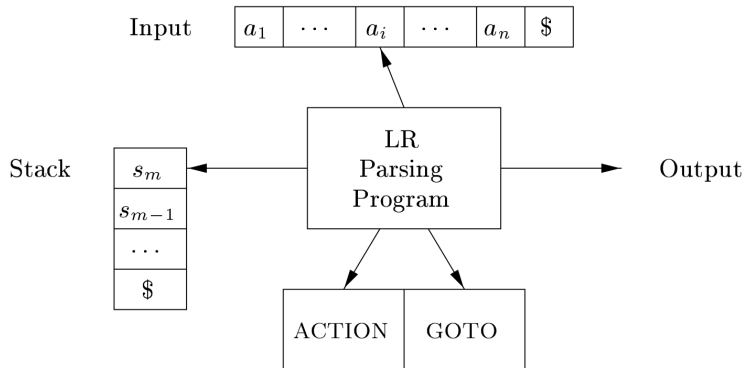


- شکل زیر روند تجزیه $id * id$ را نشان می‌دهد.

LINE	STACK	SYMBOLS	INPUT	ACTION
(1)	0	\$	id * id \$	shift to 5
(2)	0 5	\$ id	* id \$	reduce by $F \rightarrow \mathbf{id}$
(3)	0 3	\$ F	* id \$	reduce by $T \rightarrow F$
(4)	0 2	\$ T	* id \$	shift to 7
(5)	0 2 7	\$ $T *$	id \$	shift to 5
(6)	0 2 7 5	\$ $T * \mathbf{id}$	\$	reduce by $F \rightarrow \mathbf{id}$
(7)	0 2 7 10	\$ $T * F$	\$	reduce by $T \rightarrow T * F$
(8)	0 2	\$ T	\$	reduce by $E \rightarrow T$
(9)	0 1	\$ E	\$	accept

الگوریتم تجزیه LR

- شمای کلی یک تجزیه‌کننده LR در شکل زیر نمایش داده شده است.



- این تجزیه‌کننده از یک ورودی، یک خروجی، یک پشته، یک برنامه تجزیه‌کننده، و یک جدول تجزیه استفاده می‌کند که این جدول از دو بخش Action و Goto تشکیل شده است.
- برنامه تجزیه‌کننده برای همه تجزیه‌کننده‌های پایین به بالا یکسان است. تنها جدول تجزیه به ازای هر تجزیه‌کننده متفاوت خواهد بود.
- برنامه تجزیه‌کننده کاراکترها را یک‌به‌یک از ورودی می‌خواند. جایی که یک تجزیه‌کننده انتقال کاهش یک انتقال انجام می‌دهد، تجزیه‌کننده LR از یک حالت به حالت دیگر گذار می‌کند. هر حالت در واقع اطلاعات حالت‌های پشته که در زیر آن قرار دارند را خلاصه می‌کند.

- پشته شامل حالت‌های $s_0 s_1 \dots s_m$ می‌شود به طوری که s_m روی پشته است. پشته حالت‌های ماشین $LR(0)$ را نگه می‌دارد.
- هر حالت یک نماد گرامر متناظر با آن دارد. حالت‌ها متناظر هستند با مجموعه آیتم‌ها و یک گذار از حالت i به حالت j وجود دارد اگر $Goto(I_i, X) = I_j$.
- همه گذارها به حالت j برای نماد X هستند. بنابراین هر حالت به جز حالت 0 یک نماد متناظر با آن دارد.

ساختار جدول تجزیه LR

- جدول تجزیه از دو بخش تشکیل شده است : تابع Action و تابع Goto.
- ۱. تابع Action حالت i و یک ترمینال a (یا نماد $\$$ که در پایان رشته است) را دریافت می‌کند. مقدار $Action[i, a]$ یکی از چهار مورد زیر می‌تواند باشد.
 - انتقال به حالت j : حالت j که نماینده ورودی a است، وارد پشته می‌شود.
 - کاهش $A \rightarrow \beta$: جمله β که بر روی پشته قرار دارد با A کاهش پیدا می‌کند.
 - پذیرش : رشته پذیرفته می‌شود و تجزیه به اتمام می‌رسد.
 - خطا : تجزیه‌کننده یک خطا در ورودی می‌یابد و عملیات بازایی خطا انجام می‌دهد.
- ۲. تابع Goto که بر روی مجموعه‌های آیت‌ها تعریف شده بود را به حالت‌ها تعمیم می‌دهیم. اگر $Goto[I_i, A] = I_j$ ، آنگاه $Goto$ حالت i و متغیر A را به حالت j نگاشت می‌کند.

پیکربندی تجزیه کننده LR

- برای توصیف رفتار تجزیه کننده LR از یک نشانه گذاری برای نشان دادن وضعیت تجزیه کننده استفاده می کنیم، یعنی وضعیت پشته و رشته باقیمانده برای تجزیه.
- یک پیکربندی¹ از تجزیه کننده LR یک دوتایی به صورت $(s_0 s_1 \dots s_m, a_i a_{i+1} \dots a_n \$)$ است، به طوری که جزء اول محتوای پشته (بالای پشته در سمت راست) و جزء دوم باقیمانده رشته ورودی است.
- این پیکربندی نشان دهنده صورت جمله ای $X_1 X_2 \dots X_m a_i a_{i+1} \dots a_n$ است.
- درواقع X_i نماد گرامری است که با حالت s_i نمایش داده می شود.
- توجه کنید که حالت آغازین s_0 در تجزیه کننده نماینده هیچ نمادی از گرامر نیست و تنها زیر پشته را نشان می دهد.

¹ configuration

- حرکت بعدی تجزیه کننده از یک پیکربندی با خواندن نماد ورودی a_i و حالت s_m بر روی پشته توسط $Action[s_m, a_i]$ از جدول تجزیه تعیین می شود.

۱. اگر $Action[s_m, a_i] = \text{shift } s$ باشد، آنگاه تجزیه کننده یک عملیات انتقال انجام می دهد و حالت بعدی s را به پشته منتقل می کند و در پیکربندی $(s_0 s_1 \dots s_m s, a_{i+1} \dots a_n \$)$ قرار می گیرد. نیازی نیست نماد a_i بر روی پشته باشد زیرا اگر نیاز به آن بود می توان توسط حالت s آن را بازیابی کرد (البته در عمل هیچگاه نیازی به آن نیست). نماد بعدی a_{i+1} خواهد بود.

۲. اگر $\text{Action}[s_m, a_i] = \text{reduce } A \rightarrow \beta$ باشد، آنگاه تجزیه کننده یک کاهش انجام می دهد و وارد پیکربندی $(s_0 s_1 \dots s_{m-r} s, a_i a_{i+1} \dots a_n \$)$ می شود. مقدار r طول β است و $s = \text{Goto}[s_{m-r}, A]$ در اینجا تجزیه کننده ابتدا تعداد r حالت را از پشت به برمی دارد و حالت s_{m-r} بر روی پشت قرار می گیرد. سپس حالت s یعنی $\text{Goto}[s_{m-r}, A]$ بر روی پشت قرار می گیرد. دنباله نمادهای $X_{m-r+1} \dots X_m$ متناظر است با حالت های برداشته شده از روی پشت که بر β یعنی بدنه قانون کاهش منطبق می شود.
۳. اگر $\text{Action}[s_m, a_i] = \text{accept}$ آنگاه تجزیه به پایان می رسد.
۴. اگر $\text{Action}[s_m, a_i] = \text{error}$ آنگاه تجزیه کننده یک تابع بازیابی کننده خطا فراخوانی می کند.

الگوریتم تجزیه LR

- الگوریتم تجزیه‌کننده LR به صورت زیر عمل می‌کند. همهٔ تجزیه‌کننده‌های LR به همین صورت عمل می‌کنند و تنها تفاوت آنها اطلاعات ذخیره شده در Action و Goto در جدول تجزیه است.
- فرض کنید رشته w به یک تجزیه‌کننده LR برای گرامر G داده شده است.
- اگر رشته w متعلق به $L(G)$ باشد، دنباله‌ای از قوانین کاهش به صورت پایین به بالا برای رشته w به دست می‌آید در غیراینصورت پیام خطا صادر می‌شود.

الگوریتم تجزیه LR

- در ابتدا تجزیه‌کننده s_0 را بر روی پشته قرار می‌دهد و $w\$$ بر روی بافر ورودی قرار می‌گیرد. سپس الگوریتم زیر اجرا می‌شود.

```
let  $a$  be the first symbol of  $w\$$ ;  
while(1) {  
    let  $s$  be the state on top of the stack;  
    if ( ACTION[ $s, a$ ] = shift  $t$  ) {  
        push  $t$  onto the stack;  
        let  $a$  be the next input symbol;  
    } else if ( ACTION[ $s, a$ ] = reduce  $A \rightarrow \beta$  ) {  
        pop  $|\beta|$  symbols off the stack;  
        let state  $t$  now be on top of the stack;  
        push GOTO[ $t, A$ ] onto the stack;  
        output the production  $A \rightarrow \beta$ ;  
    } else if ( ACTION[ $s, a$ ] = accept ) break;  
    else call error-recovery routine;  
}
```

- گرامر زیر را در نظر بگیرید.

$$(1) \quad E \rightarrow E + T$$

$$(2) \quad E \rightarrow T$$

$$(3) \quad T \rightarrow T * F$$

$$(4) \quad T \rightarrow F$$

$$(5) \quad F \rightarrow (E)$$

$$(6) \quad F \rightarrow \mathbf{id}$$

الگوریتم تجزیه LR

- شکل زیر توابع Action و Goto از تجزیه کننده LR برای این گرامر نشان می دهد.

STATE	ACTION						GOTO		
	id	+	*	()	\$	<i>E</i>	<i>T</i>	<i>F</i>
0	s5			s4			1	2	3
1		s6				acc			
2		r2	s7		r2	r2			
3		r4	r4		r4	r4			
4	s5			s4			8	2	3
5		r6	r6		r6	r6			
6	s5			s4				9	3
7	s5			s4					10
8		s6			s11				
9		r1	s7		r1	r1			
10		r3	r3		r3	r3			
11		r5	r5		r5	r5			

- در این جدول s i به معنی انتقال به حالت i است، r j به معنی کاهش با قانون شماره j است، acc به معنی پذیرش و خانه‌های خالی به معنی خطا است.

الگوریتم تجزیه LR

- برای ورودی $id * id + id$ دنباله محتوای پشته و ورودی در شکل زیر نشان داده شده است.

	STACK	SYMBOLS	INPUT	ACTION
(1)	0		id * id + id \$	shift
(2)	0 5	id	* id + id \$	reduce by $F \rightarrow id$
(3)	0 3	F	* id + id \$	reduce by $T \rightarrow F$
(4)	0 2	T	* id + id \$	shift
(5)	0 2 7	$T *$	id + id \$	shift
(6)	0 2 7 5	$T * id$	+ id \$	reduce by $F \rightarrow id$
(7)	0 2 7 10	$T * F$	+ id \$	reduce by $T \rightarrow T * F$
(8)	0 2	T	+ id \$	reduce by $E \rightarrow T$
(9)	0 1	E	+ id \$	shift
(10)	0 1 6	$E +$	id \$	shift
(11)	0 1 6 5	$E + id$	\$	reduce by $F \rightarrow id$
(12)	0 1 6 3	$E + F$	\$	reduce by $T \rightarrow F$
(13)	0 1 6 9	$E + T$	\$	reduce by $E \rightarrow E + T$
(14)	0 1	E	\$	accept

- برای مثال در خط (۱) تجزیه‌کننده در حالت 0 است و اولین نماد id است. عملیات $s5$ باید انجام شود، بدین معنی که یک انتقال با وارد کردن حالت 5 به پشته انجام می‌شود. سپس $*$ نماد بعدی است و عملیات حالت 5 بر روی ورودی $*$ یک کاهش با $id \rightarrow F$ است. یک حالت از پشته برداشته می‌شود. چون $Goto$ در حالت 0 بر روی F حالت 3 است، پس حالت 3 بر روی پشته اضافه می‌شود.

ساختن جدول تجزیه LR

- برای استفاده از تجزیه‌کننده LR ساده یا SLR ابتدا باید جدول تجزیه آن را بسازیم.
- الگوریتم SLR با آیتم‌های $LR(0)$ و ماشین $LR(0)$ آغاز می‌کند.
- به ازای گرامر دلخواه G گرامر افزوده شده G' با متغیر شروع جدید S' ساخته می‌شود. با استفاده از G' گروه استاندارد مجموعه آیتم‌های C با تابع Goto ساخته می‌شود.
- سپس جدول تجزیه با استفاده از جدول تجزیه زیر ساخته می‌شود. قبل از ساختن جدول نیاز داریم برای همه متغیرهای A مقدار $Follow(A)$ را محاسبه کنیم.

ساختن جدول تجزیه LR

- الگوریتم ساخت جدول تجزیه SLR به صورت زیر است :

۱. گروهی از مجموعه‌های آیتم‌های $LR(0)$ برای گرامر G' به صورت $C = \{I_0, I_1, \dots, I_n\}$ می‌سازیم.
 ۲. حالت i را به ازای I_i می‌سازیم و عملیات تجزیه برای حالت i را به صورت زیر تعیین می‌کنیم :
- اگر $[A \rightarrow \alpha \cdot a\beta]$ در I_i باشد و $Goto(I_i, a) = I_j$ باشد، آنگاه $Action[i, a] = shift\ j$. در اینجا a باید یک ترمینال باشد.
 - اگر $[A \rightarrow \alpha \cdot]$ در I_i باشد، آنگاه $Action[i, a] = reduce\ A \rightarrow \alpha$ به ازای هر a در $Follow(A)$. در اینجا A نمی‌تواند S' باشد.
 - اگر $[S' \rightarrow S \cdot]$ در I_i باشد آنگاه $Action[i, \$] = accept$.

ساختن جدول تجزیه LR

- اگر در حین اجرای این الگوریتم تعارضی در عملیات به وجود آمد، می‌گوییم گرامر $SLR(1)$ نیست و الگوریتم نمی‌تواند تجزیه‌کننده تولید کند.

۳. گذار Goto برای هر حالت i برای همه متغیرهای A با استفاده از این قانون محاسبه می‌شود: اگر $Goto(I_i, A) = I_j$ آنگاه $Goto[i, A] = j$.

۴. هر خانه‌ای در جدول که برای آن در گام‌ها ۲ و ۳ مقداری تولید نشده است، خطا محسوب می‌شود.

۵. حالت اولیه تجزیه‌کننده حالتی است که از مجموعه آیتم‌هایی ساخته شده است که حاوی $[S' \rightarrow \cdot S]$ است.

- جدول تجزیه‌ای که با استفاده از این الگوریتم به دست می‌آید، جدول $SLR(1)$ برای گرامر G نامیده می‌شود. تجزیه‌کننده LR که از جدول $SLR(1)$ برای گرامر G استفاده می‌کند تجزیه‌کننده $SLR(1)$ برای G نامیده می‌شود. گرامری که برای آن یک تجزیه‌کننده $SLR(1)$ وجود داشته باشد، گرامر $SLR(1)$ نامیده می‌شود. معمولاً (۱) را حذف می‌کنیم و تجزیه‌کننده و گرامر را SLR می‌نامیم.

ساختن جدول تجزیه LR

- می‌خواهیم جدول تجزیه SLR برای گرامر زیر بسازیم.

$$(1) \quad E \rightarrow E + T$$

$$(2) \quad E \rightarrow T$$

$$(3) \quad T \rightarrow T * F$$

$$(4) \quad T \rightarrow F$$

$$(5) \quad F \rightarrow (E)$$

$$(6) \quad F \rightarrow \mathbf{id}$$

- مجموعه آیتم‌های I_0 به صورت زیر است.

$$E' \rightarrow \cdot E$$

$$E \rightarrow \cdot E + T$$

$$E \rightarrow \cdot T$$

$$T \rightarrow \cdot T * F$$

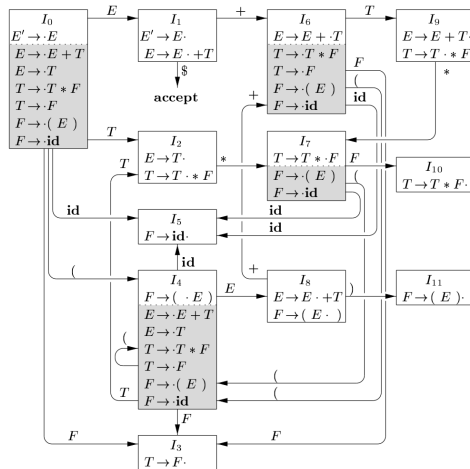
$$T \rightarrow \cdot F$$

$$F \rightarrow \cdot (E)$$

$$F \rightarrow \cdot \mathbf{id}$$

ساختن جدول تجزیه LR

- مجموعه آیت‌ها را به صورت زیر قبلاً محاسبه کردیم.



ساختن جدول تجزیه LR

- با استفاده از آیت $F \rightarrow \cdot (E)$ می‌توان مقدار $\text{Action}[0, (] = \text{shift } 4$ را محاسبه کرد و با استفاده از آیت $F \rightarrow \cdot \text{id}$ مقدار $\text{Action}[0, \text{id}] = \text{shift } 5$ به دست می‌آید. بقیه آیت‌ها در I_0 مقداری به دست نمی‌دهند.
 - حال آیت‌های I_1 را در نظر می‌گیریم. برای $E' \rightarrow E \cdot$ به دست می‌آوریم $\text{Action}[1, \$] = \text{accept}$ و برای $E \rightarrow E \cdot +$ به دست می‌آوریم $\text{Action}[1, +] = \text{shift } 6$.
 - برای آیت‌های I_2 نیز عملیات را محاسبه می‌کنیم. برای آیت $E \rightarrow T \cdot$ از آنجایی که $\text{Follow}(E) = \{ \$, +,) \}$ داریم :
- $\text{Action}[2, \$] = \text{Action}[2, +] = \text{Action}[2,)] = \text{reduce } E \rightarrow T$
- همچنین برای آیت $T \rightarrow T \cdot * F$ در I_2 داریم $\text{Action}[2, *] = \text{shift } 7$.

ساختن جدول تجزیه LR

- با ادامه این روند جدول تجزیه به صورت زیر محاسبه می‌شود.

STATE	ACTION						GOTO		
	id	+	*	()	\$	<i>E</i>	<i>T</i>	<i>F</i>
0	s5			s4			1	2	3
1		s6				acc			
2		r2	s7		r2	r2			
3		r4	r4		r4	r4			
4	s5			s4			8	2	3
5		r6	r6		r6	r6			
6	s5			s4				9	3
7	s5			s4					10
8		s6			s11				
9		r1	s7		r1	r1			
10		r3	r3		r3	r3			
11		r5	r5		r5	r5			

ساختن جدول تجزیه LR

- هر گرامر $SLR(1)$ غیر مبهم است، اما بسیاری از گرامرهای غیر مبهم وجود دارند که $SLR(1)$ نیستند.
- گرامر زیر را در نظر بگیرید.

$$\begin{array}{lcl} S & \rightarrow & L = R \mid R \\ L & \rightarrow & *R \mid \mathbf{id} \\ R & \rightarrow & L \end{array}$$

ساختن جدول تجزیه LR

- گروه استاندارد مجموعه‌های آیتم‌های $LR(0)$ برای این گرامر به صورت زیر است.

$$\begin{aligned} I_0: \quad & S' \rightarrow \cdot S \\ & S \rightarrow \cdot L = R \\ & S \rightarrow \cdot R \\ & L \rightarrow \cdot * R \\ & L \rightarrow \cdot \mathbf{id} \\ & R \rightarrow \cdot L \end{aligned}$$

$$I_5: \quad L \rightarrow \mathbf{id} \cdot$$

$$\begin{aligned} I_6: \quad & S \rightarrow L = \cdot R \\ & R \rightarrow \cdot L \\ & L \rightarrow \cdot * R \\ & L \rightarrow \cdot \mathbf{id} \end{aligned}$$

$$I_1: \quad S' \rightarrow S \cdot$$

$$I_7: \quad L \rightarrow * R \cdot$$

$$\begin{aligned} I_2: \quad & S \rightarrow L \cdot = R \\ & R \rightarrow L \cdot \end{aligned}$$

$$I_8: \quad R \rightarrow L \cdot$$

$$I_3: \quad S \rightarrow R \cdot$$

$$I_9: \quad S \rightarrow L = R \cdot$$

$$\begin{aligned} I_4: \quad & L \rightarrow * \cdot R \\ & R \rightarrow \cdot L \\ & L \rightarrow \cdot * R \\ & L \rightarrow \cdot \mathbf{id} \end{aligned}$$

ساختن جدول تجزیه LR

- مجموعه آیت‌های I_2 را در نظر بگیرید. با استفاده از آیت‌م اول به دست می‌آوریم $Action[2, =] = shift\ 6$.
از آنجایی که $Follow(R)$ حاوی $=$ است (برای مثال در اشتقاق $S \Rightarrow L = R \Rightarrow R = R$), آیت‌م دوم به دست می‌دهد $Action[2, =] = reduce\ R \rightarrow L$. چون برای $Action[2, =]$ دو مقدار انتقال و کاهش به دست می‌آوریم، پس در حالت ۲ بر روی ورودی $=$ یک تعارض انتقال و کاهش وجود دارد.
- این گرامر مبهم نیست. دلیل این تعارض این است که تجزیه‌کننده SLR به اندازه کافی قدرتمند نیست تا بتواند بر روی این گرامر تصمیم بگیرد. تجزیه‌کننده LALR مجموعه بزرگتری از گرامرها از جمله این گرامر را می‌تواند تجزیه کند.
- توجه کنید که گرامرهای غیرمبهمی وجود دارند که تجزیه‌کننده LR برای آنها وجود ندارد. البته این گرامرها در زبان‌های برنامه‌نویسی استفاده نمی‌شوند.

- حال ببینیم چرا ماشین $LR(0)$ می تواند برای تصمیم گیری در مورد انتقال و کاهش استفاده شود.
- در ماشین $LR(0)$ برای یک گرامر، پشته، پیشوندی از صورت جمله ای راست را نگهداری می کند.
- اگر پشته α را نگهداری کند و ورودی باقیمانده x باشد آنگاه دنباله ای از کاهش ها αx را به S کاهش می دهد.

$$S \xRightarrow{*}_{rm} \alpha x$$
در واقع در فرایند اشتقاق داریم
- اما همه پیشوندهای صورت های جمله ای راست نمی توانند بر روی پشته ظاهر شوند. برای مثال فرض کنید داشته باشیم $(E) * id \xRightarrow{*}_{rm} E * id \xRightarrow{*}_{rm} E$ آنگاه در فرایند تجزیه پشته می تواند حاوی (E) ، (E) و (E) باشد اما نمی تواند حاوی $(E) *$ باشد زیرا (E) یک هندل است که باید به F کاهش پیدا کند.

پیشوندهای ماندنی

- پیشوندهایی از صورت‌های جمله‌ای راست که می‌توانند بر روی پشته در یک تجزیه‌کننده انتقال‌کاهش ظاهر شوند، پیشوندهای ماندنی¹ نامیده می‌شوند. یک پیشوند ماندنی پیشوندی از یک صورت جمله‌ای راست است که یک هندل را شامل نمی‌شود.
- تجزیه SLR بر این اصل عمل می‌کند که ماشین $LR(0)$ پیشوندهای ماندنی را تشخیص می‌دهد. می‌گوییم آیت $A \rightarrow \beta_1 \cdot \beta_2$ برای پیشوند ماندنی $\alpha\beta_1$ معتبر است اگر اشتقاقی به صورت $S' \xRightarrow{*}_{rm} \alpha A w \xRightarrow{rm} \alpha\beta_1\beta_2 w$ وجود داشته باشد. یک آیت می‌تواند برای بسیاری از پیشوندهای ماندنی معتبر باشد.
- این که $A \rightarrow \beta_1 \cdot \beta_2$ برای $\alpha\beta_1$ معتبر است به ما می‌گوید وقتی $\alpha\beta_1$ را بر روی پشته مشاهده کردیم انتقال انجام دهیم یا کاهش.
- اگر $\beta_2 \neq \epsilon$ آنگاه می‌توان نتیجه گرفت که هندل به پشته انتقال داده نشده است بنابراین باید انتقال انجام شود. اگر $\beta_2 = \epsilon$ آنگاه $A \rightarrow \beta_1$ باید هندل باشد و باید با استفاده از این قانون کاهش انجام شود.

¹ viable prefix

- البته اگر دو آیتم معتبر داشته باشیم که دو عملیات متفاوت را تعیین کنند، آنگاه تعارض به وجود می‌آید که در اینصورت بررسی کاراکترهای بعدی در ورودی ممکن است به حل تعارض کمک کنند و درواقع تجزیه‌کننده‌های قدرتمندتر چنین می‌کنند.
- به سادگی می‌توانیم مجموعه آیتم‌های معتبر را برای هر پیشوند ماندنی که بر روی پشته تجزیه‌کننده می‌تواند ظاهر شود را محاسبه کنیم.
- درواقع این یک قضیه پایه‌ای مهم در نظریه تجزیه LR است که مجموعه آیتم‌های معتبر برای پیشوند ماندنی γ دقیقاً مجموعه آیتم‌هایی است که از حالت اولیه با مسیری با برچسب γ در ماشین $LR(0)$ قابل دسترسی هستند.
- مجموعه همه آیتم‌های معتبر همه اطلاعات مهم که می‌توانند در پشته قرار بگیرند را در برمی‌گیرد. این قضیه در نظریه تجزیه‌کننده اثبات می‌شود که به اثبات آن نمی‌پردازیم.

پیشوندهای ماندنی

- گرامر زیر را در نظر بگیرید.

$$(1) \quad E \rightarrow E + T$$

$$(4) \quad T \rightarrow F$$

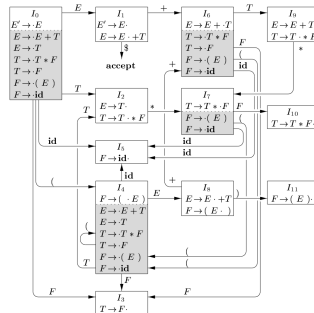
$$(2) \quad E \rightarrow T$$

$$(5) \quad F \rightarrow (E)$$

$$(3) \quad T \rightarrow T * F$$

$$(6) \quad F \rightarrow \text{id}$$

- مجموعه آیت‌های این گرامر به همراه توابع گذار را به صورت زیر قبلا محاسبه کردیم.



- رشته $E + T^*$ یک پیشوند ماندنی از گرامر است. ماشین $LR(0)$ پس از خواندن $E + T^*$ در حالت 7 قرار می‌گیرد. حالت 7 آیتم‌های زیر را شامل می‌شود که برای $E + T^*$ معتبر هستند.

$$T \rightarrow T * \cdot F$$

$$F \rightarrow \cdot (E)$$

$$F \rightarrow \cdot id$$

- برای درک دلیل این امر اشتقاق راست زیر را در نظر بگیرید.

$$\begin{array}{l}
 E' \Rightarrow E \\
 \text{rm} \\
 \Rightarrow E + T \\
 \text{rm} \\
 \Rightarrow E + T * F \\
 \text{rm}
 \end{array}$$

$$\begin{array}{l}
 E' \Rightarrow E \\
 \text{rm} \\
 \Rightarrow E + T \\
 \text{rm} \\
 \Rightarrow E + T * F \\
 \text{rm} \\
 \Rightarrow E + T * (E) \\
 \text{rm}
 \end{array}$$

$$\begin{array}{l}
 E' \Rightarrow E \\
 \text{rm} \\
 \Rightarrow E + T \\
 \text{rm} \\
 \Rightarrow E + T * F \\
 \text{rm} \\
 \Rightarrow E + T * \mathbf{id} \\
 \text{rm}
 \end{array}$$

- اشتقاق اول معتبر بودن $T \rightarrow T * F$ را نشان می‌دهد. اشتقاق دوم معتبر بودن $(E) \rightarrow F$ و اشتقاق سوم معتبر بودن $F \rightarrow id$ را نشان می‌دهد.
- می‌توان نشان داد که هیچ آیتم معتبری برای $E + T^*$ وجود ندارد.

تجزیه‌کننده‌های LR قدرتمندتر

- در این قسمت تجزیه‌کننده‌های LR را تعمیم می‌دهیم و دو تجزیه‌کننده قدرتمند را شرح می‌دهیم.

۱. تجزیه‌کننده LR استاندارد¹ یا CLR : این روش از مجموعه‌ای بزرگ از آیتم‌ها به نام آیتم‌های $LR(1)$ استفاده می‌کند.

۲. تجزیه‌کننده LR با بررسی نماد جلویی² یا LALR : این روش بر پایه مجموعه‌های آیتم‌های $LR(0)$ است و تعداد بسیار کمتری حالت نسبت به تجزیه‌کننده‌ها بر پایه $LR(1)$ دارد. تجزیه‌کننده LALR تعداد بسیار بیشتری از گرامرها را نسبت به SLR پوشش می‌دهد و جدول تجزیه‌ای که از آن استفاده می‌کند از جدول‌های SLR بزرگ‌تر نیست. در بسیاری از تجزیه‌کننده‌ها و کامپایلرها از روش LALR استفاده می‌شود.

¹ canonical LR

² lookahead LR

آیتم‌های LR(1) استاندارد

- می‌خواهیم یکی از روش‌های بسیار متداول برای تولید جدول تجزیه LR را شرح دهیم.
- در روش SLR، حالت i عملیات کاهش را با $A \rightarrow \alpha$ انجام می‌دهد اگر مجموعه آیتم‌های I_i شامل $[A \rightarrow \alpha \cdot]$ باشد و نماد ورودی a در $\text{Follow}(A)$ باشد.
- در برخی مواقع وقتی حالت i بر روی پشته است، پیشوند ماندنی $\beta\alpha$ بر روی پشته چنان است که βA نمی‌تواند با a در هیچ یک از صورت‌های جمله‌ای دنبال شود. در چنین مواردی کاهش $A \rightarrow \alpha$ بر روی ورودی a غیر معتبر است.

آیتم‌های LR(1) استاندارد

- مثال : گرامر زیر را در نظر بگیرید.

$$\begin{array}{lcl} S & \rightarrow & L = R \mid R \\ L & \rightarrow & *R \mid \mathbf{id} \\ R & \rightarrow & L \end{array}$$

آیتم‌های LR(1) استاندارد

- مجموعه آیتم‌های LR(0) را برای این گرامر به صورت زیر ساختیم.

$$\begin{aligned} I_0: \quad & S' \rightarrow \cdot S \\ & S \rightarrow \cdot L = R \\ & S \rightarrow \cdot R \\ & L \rightarrow \cdot * R \\ & L \rightarrow \cdot \mathbf{id} \\ & R \rightarrow \cdot L \end{aligned}$$

$$I_5: \quad L \rightarrow \mathbf{id} \cdot$$

$$\begin{aligned} I_6: \quad & S \rightarrow L = \cdot R \\ & R \rightarrow \cdot L \\ & L \rightarrow \cdot * R \\ & L \rightarrow \cdot \mathbf{id} \end{aligned}$$

$$I_1: \quad S' \rightarrow S \cdot$$

$$I_7: \quad L \rightarrow * R \cdot$$

$$\begin{aligned} I_2: \quad & S \rightarrow L \cdot = R \\ & R \rightarrow L \cdot \end{aligned}$$

$$I_8: \quad R \rightarrow L \cdot$$

$$I_3: \quad S \rightarrow R \cdot$$

$$I_9: \quad S \rightarrow L = R \cdot$$

$$\begin{aligned} I_4: \quad & L \rightarrow * \cdot R \\ & R \rightarrow \cdot L \\ & L \rightarrow \cdot * R \\ & L \rightarrow \cdot \mathbf{id} \end{aligned}$$

آیتم‌های $LR(1)$ استاندارد

- در حالت ۲ آیتم $R \rightarrow L \cdot$ را داشتیم. فرض کنید این قانون همان $A \rightarrow \alpha$ است و نماد a در ورودی در اینجا علامت $=$ است که در $Follow(R)$ است.
- تجزیه‌کننده SLR کاهش با استفاده از $R \rightarrow L$ را در حالت ۲ با خواندن نماد $=$ فراخوانی می‌کند.
- اما هیچ صورت جمله‌ای در این گرامر که با $R = \dots$ آغاز شود وجود ندارد. بنابراین در حالت ۲ که حالت متناظر با پیشوند ماندنی L است نباید L را با R کاهش دهد.

آیتم‌های $LR(1)$ استاندارد

- در چنین مواردی نیاز داریم اطلاعات بیشتری دریافت کنیم تا به ما کمک کند برخی از کاهش‌ها را انجام ندهیم.
- می‌توانیم در هریک از حالت‌ها تجزیه‌کننده LR مشخص کنیم کدام نمادها می‌توانند همدل α را دنبال کنند وقتی یک کاهش به صورت $A \rightarrow \alpha$ وجود داشته باشد.
- این اطلاعات اضافی را بدین صورت در جدول تجزیه درج می‌کنیم که آیتم‌ها یک ترمینال را به عنوان مؤلفه دوم شامل شوند.
- بنابراین یک آیتم به صورت $[A \rightarrow \alpha \cdot \beta, a]$ درمی‌آید که در آن $A \rightarrow \alpha\beta$ یک قانون تولید و a یک نماد الفبا یا نماد پایان رشته $\$$ است. چنین آیتم‌هایی را آیتم $LR(1)$ می‌نامیم.

آیتم‌های LR(1) استاندارد

- عدد ۱ در LR(1) طول مؤلفه دوم در آیتم است که نمادهای جلویی¹ در آیتم نامیده می‌شود.
- نماد جلویی هیچ تأثیری در آیتمی که به صورت $[A \rightarrow \alpha \cdot \beta, a]$ است ندارد اگر β رشته تهی نباشد. اما در آیتم $[A \rightarrow \alpha \cdot, a]$ کاهش با استفاده از $A \rightarrow \alpha$ تنها صورتی انجام می‌شود که نماد بعدی a باشد.
- می‌گوییم آیتم LR(1) به صورت $[A\alpha \cdot \beta, a]$ برای پیشوند ماندنی γ معتبر است اگر اشتقاقی به صورت $S \xRightarrow{*}_{rm} \delta A w \xRightarrow{rm} \delta \alpha \beta w$ وجود داشته باشد که $\gamma = \delta \alpha$ (۱) باشد و (۲) یا a اولین نماد w باشد و یا w تهی باشد و a نماد $\$$ باشد.

¹ lookahead

آیتم‌های LR(1) استاندارد

- مثال : گرامر زیر را در نظر بگیرید.

$$\begin{aligned} S &\rightarrow B B \\ B &\rightarrow a B \mid b \end{aligned}$$

- یک اشتقاق راست $S \xRightarrow{*}_{rm} aaBab \xRightarrow{rm} aaaBab$ وجود دارد. می‌بینیم که آیتم $[B \rightarrow a \cdot B, a]$ برای پیشوند ماندنی $\gamma = aaa$ معتبر است اگر $\delta = aa$ و $A = B$ و $w = ab$ و $\alpha = a$ و $\beta = B$ باشد. همچنین اشتقاق $S \xRightarrow{*}_{rm} BaB \xRightarrow{rm} BaaB$ وجود دارد. از این اشتقاق می‌بینیم که آیتم $[B \rightarrow a \cdot B, \$]$ برای پیشوند ماندنی Baa معتبر است.

ساختن مجموعه‌های آیتم‌های $LR(1)$

- روش ساختن گروه مجموعه‌های آیتم‌های $LR(1)$ معتبر شبیه ساخت مجموعه آیتم‌های $LR(0)$ است. تنها تفاوت در توابع Closure و Goto است.

ساختن مجموعه‌های آیتم‌های LR(1)

- برای گرامر G' مجموعه آیتم‌های LR(1) با استفاده از الگوریتم زیر محاسبه می‌شود.

```
SetOfItems CLOSURE( $I$ ) {  
    repeat  
        for ( each item  $[A \rightarrow \alpha \cdot B \beta, a]$  in  $I$  )  
            for ( each production  $B \rightarrow \gamma$  in  $G'$  )  
                for ( each terminal  $b$  in FIRST( $\beta a$ ) )  
                    add  $[B \rightarrow \cdot \gamma, b]$  to set  $I$ ;  
    until no more items are added to  $I$ ;  
    return  $I$ ;  
}  
  
SetOfItems GOTO( $I, X$ ) {  
    initialize  $J$  to be the empty set;  
    for ( each item  $[A \rightarrow \alpha \cdot X \beta, a]$  in  $I$  )  
        add item  $[A \rightarrow \alpha X \cdot \beta, a]$  to set  $J$ ;  
    return CLOSURE( $J$ );  
}  
  
void items( $G'$ ) {  
    initialize  $C$  to  $\{\text{CLOSURE}(\{[S' \rightarrow \cdot S, \$]\})\}$ ;  
    repeat  
        for ( each set of items  $I$  in  $C$  )  
            for ( each grammar symbol  $X$  )  
                if ( GOTO( $I, X$ ) is not empty and not in  $C$  )  
                    add GOTO( $I, X$ ) to  $C$ ;  
    until no new sets of items are added to  $C$ ;  
}
```

ساختن مجموعه‌های آیتم‌های LR(1)

- برای اینکه بفهمیم چرا b باید در $\text{First}(\beta a)$ باشد، آیتمی به صورت $[A \rightarrow \alpha \cdot B\beta, a]$ را در نظر بگیرید که در مجموعه آیتم‌های معتبر برای پیشوند ماندنی γ است. آنگاه اشتقاق راست $S \xRightarrow{*}_{rm} \delta A a x \xRightarrow{rm} \delta \alpha B \beta a x$ وجود دارد به طوری که $\gamma = \delta \alpha$.
- فرض کنید $\beta a x$ رشته by را مشتق کند. آنگاه برای هر قانون $B \rightarrow \eta$ اشتقاق $\gamma B b y \xRightarrow{rm} \gamma \eta b y$ را $S \xRightarrow{*}_{rm} \gamma B b y$ داریم. بنابراین $[B \rightarrow \cdot \eta, b]$ برای γ معتبر است.
- توجه کنید که b می‌تواند اولین ترمینال مشتق شده از β باشد یا ممکن است β رشته تهی در اشتقاق $\beta a x \xRightarrow{*}_{rm} by$ مشتق کند.
- برای خلاصه این دو احتمال، می‌گوییم b می‌تواند هر ترمینالی در $\text{First}(\beta a x)$ باشد. توجه کنید که x نمی‌تواند اولین ترمینال by را شامل شود پس $\text{First}(\beta a x) = \text{First}(\beta a)$.

ساختن مجموعه‌های آیتم‌های LR(1)

- مثال : گرامر زیر را در نظر بگیرید.

$$\begin{aligned} S' &\rightarrow S \\ S &\rightarrow C C \\ C &\rightarrow c C \mid d \end{aligned}$$

- ابتدا closur بر روی $[S' \rightarrow \cdot S, \$]$ را محاسبه می‌کنیم. آیتم $[S' \rightarrow \cdot S, \$]$ را بر $[A \rightarrow \alpha \cdot B \beta, a]$ تطبیق می‌دهیم. پس $A = S'$ ، $\alpha = \epsilon$ ، $B = S$ ، $\beta = \beta$ و $a = \$$ است. تابع Closure می‌گوید برای هر قانون $B \rightarrow \gamma$ آیتم $[B \rightarrow \cdot \gamma, b]$ را به ازای هر b در $\text{First}(\beta a)$ اضافه کنیم. در این گرامر $B \rightarrow \gamma$ باید $S \rightarrow CC$ باشد زیرا β تهی است و a نماد $\$$ است، پس b تنها می‌تواند $\$$ باشد. پس آیتم $[S \rightarrow \cdot CC, \$]$ را می‌افزاییم.

ساختن مجموعه‌های آیتم‌های LR(1)

- برای ادامه محاسبه closure همه آیتم‌های $[C \rightarrow \cdot \gamma, b]$ را برای b در $\text{First}(C\$)$ اضافه می‌کنیم. با تطبیق $[S \rightarrow \cdot CC, \$]$ بر $[A \rightarrow \alpha \cdot B\beta, a]$ خواهیم داشت $A = S$ ، $\alpha = \epsilon$ ، $B = C$ ، $\beta = C$ و $a = \$$. از آنجایی که C رشته تهی تولید نمی‌کند، $\text{First}(C\$) = \text{First}(C)$. از آنجایی که $\text{First}(C)$ ترمینال‌های c و d را شامل می‌شود، آیتم‌های $[C \rightarrow \cdot cC, c]$ و $[C \rightarrow \cdot cC, d]$ و $[C \rightarrow \cdot d, c]$ و $[C \rightarrow \cdot d, d]$ را می‌افزاییم. هیچ‌کدام از این آیتم‌ها در سمت راست نقطه متغیر ندارد، بنابراین محاسبه اولین مجموعه $LR(0)$ به اتمام می‌رسد.

$$\begin{aligned} I_0 : \quad & S \rightarrow \cdot S, \$ \\ & S \rightarrow \cdot CC, \$ \\ & C \rightarrow \cdot cC, c/d \\ & C \rightarrow \cdot d, c/d \end{aligned}$$

- در اینجا برای سادگی به جای دو آیتم $[C \rightarrow \cdot cC, c]$ و $[C \rightarrow \cdot cC, d]$ می‌نویسیم c/d و $C \rightarrow \cdot cC$.

ساختن مجموعه‌های آیتم‌های LR(1)

- حال باید $\text{Goto}(I_0, X)$ را برای مقادیر مختلف X محاسبه کنیم. اگر $X = S$ باشد آیتم $[S' \rightarrow S \cdot, \$]$ به دست می‌آید. بنابراین داریم.

$$I_1 : \quad S' \rightarrow S \cdot, \$$$

- برای $X = C$ آیتم $[S \rightarrow C \cdot C, \$]$ به وجود می‌آید. همه قوانین متغیر C را با نماد $\$$ به عنوان مؤلفه دوم می‌افزاییم و به دست می‌آوریم :

$$\begin{aligned} I_2 : \quad & S \rightarrow C \cdot C, \$ \\ & C \rightarrow \cdot cC, \$ \\ & C \rightarrow \cdot d, \$ \end{aligned}$$

ساختن مجموعه‌های آیتم‌های LR(1)

- اگر داشته باشیم $X = c$ آنگاه آیتم $[C \rightarrow c \cdot C, c/d]$ را به دست می‌آوریم. همه قوانین متغیر C را با نماد c/d به عنوان مؤلفه دوم می‌افزاییم.

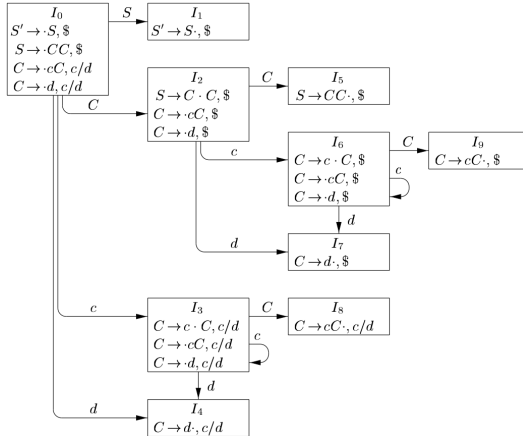
$$I_3 : \begin{array}{l} C \rightarrow c \cdot C, \ c/d \\ C \rightarrow \cdot cC, \ c/d \\ C \rightarrow \cdot d, \ c/d \end{array}$$

- در نهایت قرار می‌دهیم $X = d$ و به دست می‌آوریم.

$$I_4 : \quad C \rightarrow d \cdot, \ c/d$$

ساختن مجموعه‌های آیتم‌های LR(1)

- به همین ترتیب با محاسبه آیتم‌ها و توابع Goto گراف زیر را به دست می‌آوریم.



جداول تجزیه LR(1) استاندارد

- حال الگوریتمی را توصیف می‌کنیم که برای یک گرامر جدول تجزیه LR استاندارد با توابع Action و Goto تولید می‌کند.

۱. برای گرامر افزوده شده G' گروه مجموعه‌های آیتم‌های LR(1) را به صورت $C' = \{I_0, I_1, \dots, I_n\}$ می‌سازیم.

۲. حالت i از تجزیه‌کننده از I_i ساخته می‌شود. عملیات تجزیه برای i به صورت زیر تعیین می‌شود.

- اگر (a) $[A \rightarrow \alpha \cdot a\beta, b]$ در I_i باشد و $\text{Goto}(I_i, a) = I_j$ باشد، آنگاه $\text{Action}[i, a] = \text{shift } j$. در اینجا a باید یک ترمینال باشد.

- اگر (b) $[A \rightarrow \alpha \cdot, a]$ در I_i باشد و $A \neq S'$ ، آنگاه $\text{Action}[i, a] = \text{reduce } A \rightarrow \alpha$.

- اگر (c) $[S' \rightarrow S \cdot, \$]$ در I_i باشد، آنگاه $\text{Action}[i, \$] = \text{accept}$ اگر هر تعارض در عملیات بالا رخ دهد می‌گوییم گرامر LR(1) نیست.

جداول تجزیه LR(1) استاندارد

۳. توابع گذار Goto برای حالت i برای همه متغیرهای A به صورت زیر ساخته می‌شود: اگر $Goto(I_i, A) = I_j$ ، آنگاه $Goto[i, A] = j$.

۴. همه خانه‌هایی که در گام‌های (۲) و (۳) تعریف نشده‌اند، خطا محسوب می‌شوند.

۵. حالت اولیه تجزیه‌کننده، حالتی است که از مجموعه آیتم‌های حاوی $[S' \rightarrow \cdot S, \$]$ تشکیل شده است.

جداول تجزیه $LR(1)$ استاندارد

- جدولی که با استفاده از الگوریتم قبل تولید می‌شود جدول تجزیه $LR(1)$ استاندارد¹ نامیده می‌شود.
- یک تجزیه‌کننده LR که از چنین جدولی استفاده کند، تجزیه‌کننده $LR(1)$ استاندارد نامیده می‌شود.
- گرامری که برای آن یک تجزیه $LR(1)$ استاندارد وجود داشته باشد گرامر $LR(1)$ نامیده می‌شود.

¹ canonical $LR(1)$

جداول تجزیه LR(1) استاندارد

- گرامر زیر را در نظر بگیرید.

$$\begin{array}{lcl} S' & \rightarrow & S \\ S & \rightarrow & C C \\ C & \rightarrow & c C \mid d \end{array}$$

جداول تجزیه LR(1) استاندارد

- جدول تجزیه LR(1) استاندارد برای این گرامر در زیر نشان داده شده است. قوانین ۱، ۲ و ۳ به ترتیب $S \rightarrow CC$ ، $C \rightarrow cC$ و $C \rightarrow d$ هستند.

STATE	ACTION			GOTO	
	<i>c</i>	<i>d</i>	\$	<i>S</i>	<i>C</i>
0	s3	s4		1	2
1			acc		
2	s6	s7			5
3	s3	s4			8
4	r3	r3			
5			r1		
6	s6	s7			9
7			r3		
8	r2	r2			
9			r2		

جداول تجزیه $LR(1)$ استاندارد

- هر گرامر $SLR(1)$ یک گرامر $LR(1)$ است اما برای یک گرامر $SLR(1)$ تجزیه کننده LR استاندارد ممکن است تعداد حالت‌های بیشتری نسبت به تجزیه کننده SLR برای همان گرامر داشته باشد.

ساختن جداول تجزیه LALR

- حال به معرفی تجزیه‌کننده LALR¹ می‌پردازیم. این تجزیه‌کننده در عمل بسیار مورد استفاده قرار می‌گیرد، زیرا جداول تجزیه آن از جدول تجزیه‌کننده LR استاندارد کوچک‌تر هستند و بیشتر زبان‌های برنامه‌نویسی را می‌توان با استفاده از گرامر LALR می‌توان توصیف کرد.
- جداول SLR نیز نسبت به CLR کوچک‌تر هستند اما برخی از ساختارهای زبان‌های برنامه‌نویسی را نمی‌توان با استفاده از گرامرهای SLR توصیف کرد.
- جداول تجزیه SLR و LALR تقریباً برابرند و در زبانی مانند زبان سی حدود چند صد حالت دارند، اما جدول تجزیه CLR برای زبان سی حدود چند هزار حالت خواهد داشت.

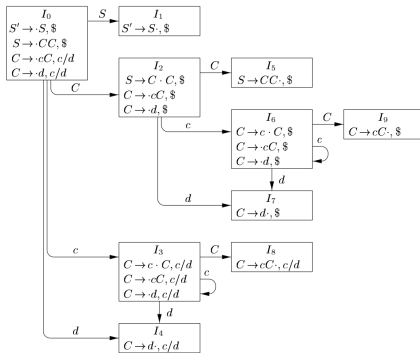
¹ lookahead LR

ساختن جداول تجزیه LALR

- گرامر زیر را در نظر بگیرید.

$$\begin{aligned} S' &\rightarrow S \\ S &\rightarrow CC \\ C &\rightarrow cC \mid d \end{aligned}$$

- مجموعه آیتم‌های LR(1) برای این گرامر در شکل زیر نشان داده شده‌اند.



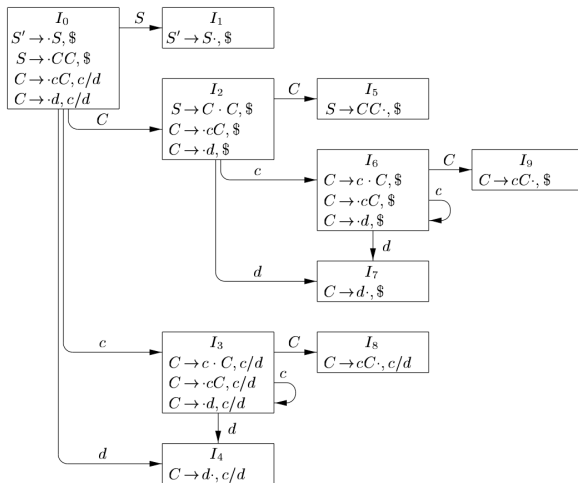
- دو حالت شبیه به هم I_4 و I_7 را در نظر بگیرید. هریک از این حالت‌ها فقط آیتم‌هایی با مؤلفه اول $C \rightarrow d \cdot$ دارند. در I_4 مؤلفه دوم c یا d است و در I_7 مؤلفه دوم $\$$ است.
- برای اینکه تفاوت I_4 و I_7 را بفهمیم، یک مثال را بررسی می‌کنیم. این گرامر زبان منظم c^*dc^*d را تولید می‌کند. وقتی ورودی $cc \dots cdcc \dots cd$ خوانده می‌شود، تجزیه‌کننده اولین گروه از c ها و کاراکتر d پس از آن را به پشت‌انتقال می‌دهد و وارد حالت ۴ می‌شود. سپس کاهش توسط $C \rightarrow d$ انجام می‌شود با توجه به اینکه کاراکتر بعدی می‌تواند c یا d باشد. اگر $\$$ به دنبال اولین d بیاید، یک ورودی به صورت ccd داریم که در زبان نیست و حالت ۴ به درستی با خواندن $\$$ وجود خطا را نشان می‌دهد.

ساختن جداول تجزیه LALR

- تجزیه‌کننده بعد از خواندن دومین d وارد حالت ۷ می‌شود. پس از آن باید در ورودی $\$$ خوانده شود وگرنه ورودی برطبق الگوی s^*dc^*d نیست. بنابراین حالت ۷ و ورودی $\$$ کاهش $C \rightarrow d$ انجام می‌شود و با ورودی c یا d خطا صادر می‌شود.
- حال فرض کنید I_4 و I_7 را با I_{47} جایگزین کنیم که اجتماع I_4 و I_7 است و از سه آیت $[C \rightarrow d, c/d/\$]$ تشکیل شده است.
- عملیات حالت 47 اکنون این است که بر روی کاهش انجام می‌دهد در حالی که قبل از ادغام دو حالت برخی از شرایط منجر به خطا می‌شدند. البته خطا اکنون نیز تشخیص داده خواهد شد.
- در حالت کلی می‌توانیم آیتم‌هایی را که هسته یکسان دارند یا به عبارت دیگر مؤلفه اول آنها یکسان است ادغام کنیم.

ساختن جداول تجزیه LALR

- آیتم‌های زیر را در نظر بگیرید.



- برای مثال در I_4 و I_7 آیتم‌هایی با هسته $\{C \rightarrow d \cdot\}$ وجود دارد. همچنین در I_3 و I_6 هسته $\{C \rightarrow c \cdot C, C \rightarrow \cdot cC, C \rightarrow \cdot d\}$ وجود دارد. در I_8 و I_9 نیز هسته $C \rightarrow cC \cdot$ وجود دارد.
- یک هسته یک مجموعه از آیتم‌های $LR(0)$ برای یک گرامر است و یک گرامر $LR(1)$ ممکن است بیش از دو مجموعه از آیتم‌ها با یک هسته تولید کند.

ساختن جداول تجزیه LALR

- از آنجایی که هسته $Goto(I, X)$ فقط به هسته I بستگی دارد، توابع $Goto$ از مجموعه‌های ادغام شده نیز ادغام می‌شوند.
- فرض کنید یک گرامر $LR(1)$ داریم. اگر همهٔ حالت‌ها با هستهٔ یکسان را ادغام کنیم، این احتمال وجود دارد که نتیجه دارای تعارض باشد اما به احتمال زیاد تعارض رخ نخواهد داد.
- می‌توان اثبات کرد که حاصل ادغام هیچ‌گاه تعارض انتقال‌کاهش نخواهد داشت، اما این امکان وجود دارد که تعارض کاهش‌کاهش رخ دهد.

ساختن جداول تجزیه LALR

- مثال : گرامر زیر را در نظر بگیرید.

$$\begin{aligned} S' &\rightarrow S \\ S &\rightarrow a A d \mid b B d \mid a B e \mid b A e \\ A &\rightarrow c \\ B &\rightarrow c \end{aligned}$$

- این گرامر چهار رشته acd و ace و bcd و bce را تولید می‌کند. این گرامر یک گرامر $LR(1)$ است.

- مجموعه آیتم‌های $\{[A \rightarrow c\cdot, d], [B \rightarrow c\cdot, e]\}$ برای پیشوند ماندنی ac معتبر است و $\{[A \rightarrow c\cdot, e], [B \rightarrow c\cdot, d]\}$ برای bc معتبر است. هیچ‌کدام از این مجموعه‌ها تعارض ندارند. اما اجتماع آنها تعارض کاهش کاهش ایجاد می‌کند.

$$\begin{aligned} A &\rightarrow c\cdot, d/e \\ B &\rightarrow c\cdot, d/e \end{aligned}$$

- برای ساخت جدول LALR ابتدا مجموعه آیت‌های $LR(1)$ را می‌سازیم و اگر تعارض ایجاد نشود هسته‌های یکسان را ادغام می‌کنیم. سپس یک جدول تجزیه از آیت‌های ادغام شده می‌سازیم. اگر امکاه چنین جدولی وجود داشت گرامر $LALR(1)$ است.

ساختن جداول تجزیه LALR

- یک الگوریتم ساده برای ساخت جدول تجزیه LALR به شرح زیر است.

۱. گروه مجموعه آیت‌های $LR(1)$ را به صورت $C' = \{I_0, I_1, \dots, I_n\}$ می‌سازیم.

۲. برای هر هسته در بین مجموعه آیت‌های $LR(1)$ مجموعه‌هایی هسته یکسان دارند را پیدا کرده مجموعه‌های آنها را ادغام می‌کنیم.

۳. فرض کنیم مجموعه‌های آیت‌های $LR(1)$ به دست آمده $C' = \{J_0, J_1, \dots, J_m\}$ است. عملیات برای حالت i از آیت J_i به همان صورتی که قبلاً توضیح داده شده به دست می‌آید. اگر یک تعارض وجود داشته باشد، تجزیه‌کننده $LALR(1)$ نیست.

۴. تابع $Goto$ به صورت زیر محاسبه می‌شود. اگر J اجتماع یک یا مجموعه آیت‌های $LR(1)$ باشد یعنی $J = I_1 \cup I_2 \cup \dots \cup I_k$ آنگاه هسته‌ها $Goto(I_1, X)$ و $Goto(I_2, X)$ و \dots و $Goto(I_k, X)$ یکسان هستند زیر I_1, I_2, \dots, I_k هسته یکسان دارند. فرض کنید k اجتماع همه مجموعه‌های آیت‌هایی باشد که هسته آنها $Goto(I_1, X)$ است. آنگاه $Goto(J, X)$.

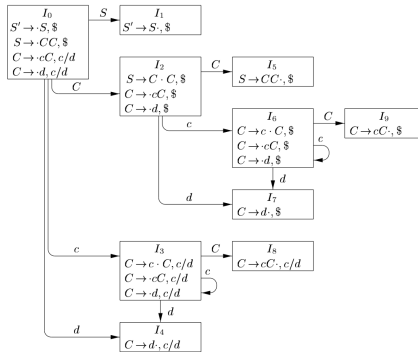
- جدولی که از الگوریتم قبل به دست می‌آید جدول تجزیه LALR نامیده می‌شود و اگر تعارض وجود نداشته باشد گرامر به دست آمده $LALR(1)$ نامیده می‌شود.

ساختن جداول تجزیه LALR

- مثال : گرامر زیر را در نظر بگیرید.

$$\begin{aligned} S' &\rightarrow S \\ S &\rightarrow C C \\ C &\rightarrow c C \mid d \end{aligned}$$

- گراف توابع Goto برای این گرامر در زیر نشان داده شده است.



- در مجموعه آیتم‌ها می‌توانیم سه جفت از آیتم‌ها را ادغام کنیم.
- مجموعه آیتم‌های I_3 و I_6 به صورت زیر می‌توانند ادغام شوند.
 $I_{36}:$ $C \rightarrow c \cdot C, c/d/\$$
 $C \rightarrow \cdot cC, c/d/\$$
 $C \rightarrow \cdot d, c/d/\$$
- مجموعه آیتم‌های I_4 و I_7 را می‌توانیم به صورت زیر ادغام کنیم.
 $I_{47}: C \rightarrow d \cdot, c/d/\$$
- و همچنین مجموعه آیتم‌های I_8 و I_9 را می‌توانیم ادغام کنیم.
 $I_{89}: C \rightarrow cC \cdot, c/d/\$$

ساختن جداول تجزیه LALR

- جدول LALR به صورت زیر به دست خواهد آمد.

STATE	ACTION			GOTO	
	<i>c</i>	<i>d</i>	\$	<i>S</i>	<i>C</i>
0	s36	s47		1	2
1			acc		
2	s36	s47			5
36	s36	s47			89
47	r3	r3	r3		
5			r1		
89	r2	r2	r2		

- تابع $Goto(I_{36}, C) = I_8$ را در نظر بگیرید. در مجموعه آیت‌های $LR(1)$ داریم $Goto(I_3, C) = I_8$ و اکنون عضوی از I_{89} است، بنابراین $Goto(I_{36}, C) = I_{89}$. از طرف دیگر اگر I_6 را در نظر بگیریم به همین نتیجه می‌رسیم، زیرا $Goto(I_6, C) = I_9$ و نیز عضوی از I_{89} است.

ساختن جداول تجزیه LALR

- زبان c^*dc^*d را بار دیگر در نظر بگیرید. تجزیه‌کننده LR و تجزیه‌کننده LALR برای این زبان شبیه به یکدیگر عمل می‌کنند و دنباله عملیات انتقال و کاهش مشابه انجام می‌دهند.

STATE	ACTION			GOTO	
	<i>c</i>	<i>d</i>	\$	<i>S</i>	<i>C</i>
0	s36	s47		1	2
1			acc		
2	s36	s47			5
36	s36	s47			89
47	r3	r3	r3		
5			r1		
89	r2	r2	r2		

STATE	ACTION			GOTO	
	<i>c</i>	<i>d</i>	\$	<i>S</i>	<i>C</i>
0	s3	s4		1	2
1			acc		
2	s6	s7			5
3	s3	s4			8
4	r3	r3			
5			r1		
6	s6	s7			9
7			r3		
8	r2	r2			
9			r2		

ساختن جداول تجزیه LALR

- برای مثال، اگر تجزیه‌کننده LR آیت‌های I_3 و I_6 را بر روی پشته قرار دهد، تجزیه‌کننده LALR نیز حالت I_{36} را بر روی پشته قرار می‌دهد.
- در حالت کلی هر تجزیه‌کننده LR و LALR معادل آن برای ورودی‌های درست عملیات مشابه انجام می‌دهند.
- وقتی در ورودی خطا وجود داشته باشد، تجزیه‌کننده LALR ممکن است تعداد بیشتری کاهش انجام دهد تا به خطا برسد، اما تجزیه‌کننده LALR هیچ‌گاه عملیات انتقال پس از رسیدن به نقطه خطای تجزیه‌کننده LR انجام نمی‌دهد.

- برای مثال، برای ورودی $\$ccd$ ، تجزیه‌کننده LR حالات 0 3 3 4 را بر روی پشته قرار می‌دهد و در حالت ۴ یک خطا تشخیص می‌دهد. تجزیه‌کننده LALR حالات 0 36 36 47 را بر روی پشته قرار می‌دهد، اما در حالت ۴۷ با ورودی $\$$ عملیات کاهش $d \rightarrow C$ را انجام می‌دهد و بر روی پشته 0 36 36 89 قرار می‌گیرد. سپس یک عملیات کاهش دیگر با استفاده از $cC \rightarrow C$ انجام داده و 0 36 89 بر روی پشته قرار می‌گیرد و در نهایت با یک کاهش دیگر 0 2 بر روی پشته قرار گرفته می‌شود. در نهایت در حالت ۲ با ورودی $\$$ تجزیه‌کننده خطا صادر می‌کند.

ساختن جداول تجزیه LALR

- الگوریتم سریع‌تری برای ساخت جدول تجزیه LALR وجود دارد که در اینجا به آن نمی‌پردازیم.
- یک زبان برنامه‌نویسی معمول با ۵۰ تا ۱۰۰ ترمینال و حدود ۱۰۰ قانون تولید می‌تواند یک جدول تجزیه با چند صد حالت تولید کند. بسیاری از خانه‌ها در جدول تجزیه تکراری هستند و بنابراین روش‌هایی برای فشرده‌سازی جدول تجزیه وجود دارد.

- معمولا برای گرامرهای مبهم نمی‌توان تجزیه‌کننده LR تولید کرد، اما برای برخی از گرامرهای مبهم می‌توان جدول تجزیه را به نحوی طراحی کرد که همیشه تصمیم درست در فرایند تجزیه اتخاذ کرد و تنها یک درخت تجزیه تولید کند. در برخی مواقع گرامر مبهم برای توصیف زبان ساده‌تر از معادل غیرمبهم آن است و بنابراین گاه می‌توان در شرایط خاص از گرامرهای مبهم در تجزیه‌کننده‌های LR استفاده کرد.

بازیابی خطا در تجزیه‌کننده LR

- تجزیه‌کننده LR خطاها را با مراجعه به جدول تجزیه شناسایی می‌کند.
- تجزیه‌کننده LR استاندارد به محض وقوع خطا، آن را شناسایی می‌کند، اما تجزیه‌کننده‌های SLR و LALR ممکن است قبل از صدور خطا تعدادی عملیات کاهش انجام دهند.
- در تجزیه‌کننده LR بازیابی خطا با توکن همگام‌کننده¹ به صورت زیر انجام می‌شود.
- پشته بررسی می‌شود تا به حالت s برسیم که با تابع `goto` به یک متغیر A گذار می‌کند. سپس از تعداد صفر یا بیشتر نمادهای ورودی چشم‌پوشی می‌شود تا اینکه به نماد a برسیم که می‌تواند متغیر A را دنبال کند. سپس حالت $Goto(s, A)$ بروی پشته قرار می‌گیرد و تجزیه ادامه پیدا می‌کند.

¹ panic-mode error recovery

- ممکن است چند انتخاب برای متغیر A وجود داشته باشد. معمولاً متغیری انتخاب می‌شود که نماینده یک قطعه از برنامه باشد برای مثال یک بلوک یا یک عبارت. برای مثال اگر A متغیر `stmt` باشد آنگاه نماد a می‌تواند نقطه‌ویرگول یا آکولاد بسته باشد که پایان دستور را مشخص می‌کند.
- این روش بازیابی خطا سعی می‌کند عبارتی را که شامل خطای نحوی است حذف کند. تجزیه‌کننده تشخیص می‌دهد که رشته‌ای که از متغیر A به دست می‌آید دارای خطا است. قسمتی از آن رشته پردازش شده است و نتیجه این پردازش تعدادی حالت بر روی پشته است. مابقی زیررشته دارای خطا در ورودی است و تجزیه‌کننده سعی می‌کند از قسمتی از ورودی چشم‌پوشی کند تا به کاراکتری برسد که متغیر A را دنبال می‌کند. با حذف تعدادی حالت از روی پشته و چشم‌پوشی از قسمتی از ورودی و قرار دادن $Goto(s, A)$ بر روی پشته، تجزیه‌کننده به احتمال زیاد می‌تواند قسمتی از ورودی که دارای خطاست را پشت سر بگذارد و با تجزیه برنامه به صورت عادی ادامه می‌دهد.

- بازیابی خطا با جایگزینی توکن‌ها¹ بدین صورت پیاده‌سازی می‌شود که هریک از خانه‌های خطا در جدول تجزیه بررسی شده و تشخیص داده می‌شود چه نوع خطاهای رایج برنامه‌نویسی ممکن است در آن مواقع رخ دهد. برای هریک از خطاها در پیام خطای مناسب تهیه می‌شود و حالت‌هایی که باید از پشته حذف شوند و قسمتی از ورودی که باید از آن چشم‌پوشی شود مشخص می‌شوند.
- هریک از خانه‌های خطا در جدول تجزیه با اشاره‌گری به یک تابع مناسب جایگزین می‌شود. این توابع می‌توانند نمادهایی را در ورودی اضافه کنند و یا قسمتی از ورودی را حذف کنند یا تغییر دهند. باید اطمینان حاصل شود که پردازش خطا باعث نمی‌شود تجزیه‌کننده وارد یک حلقه بی‌پایان شود.

¹ phrase-level error recovery