PHENIKAA UNIVERSITY
Realizing your potential

# CHƯƠNG TRÌNH DỊCH

## Chương 3. Phân tích từ vựng

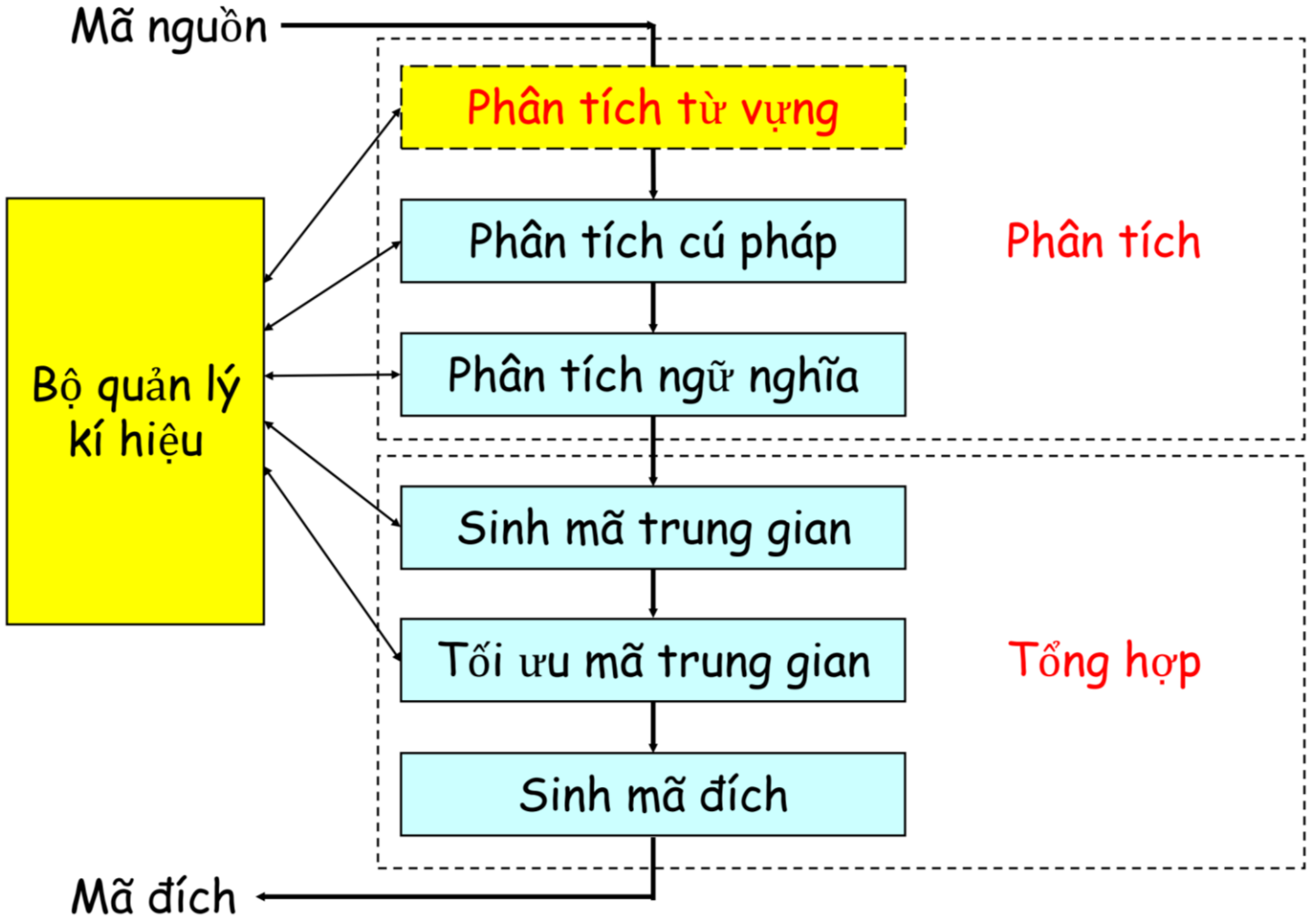
**TS. Phạm Văn Cảnh**
**Khoa Công nghệ thông tin**

Email: canh.phamvan@phenikaa-uni.edu.vn

# Nội dung

1. Vai trò của phân tích từ vựng
2. Nhiệm vụ của phân tích từ vựng
3. Các bước trong phân tích từ vựng

# Cấu trúc chương trình dịch



# 1. Vai trò của phân tích từ vựng

PHENIKAA UNIVERSITY
Realizing your potential

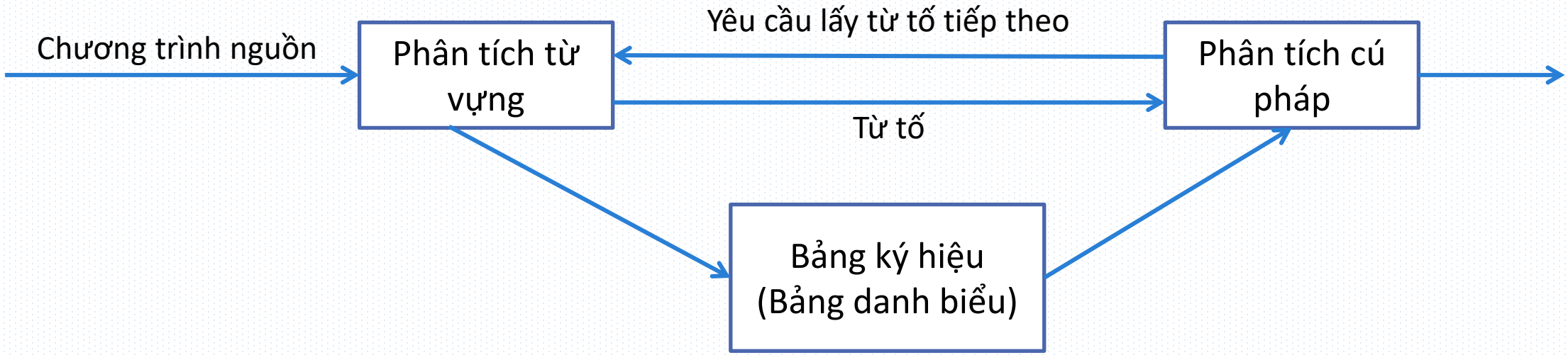
- Phân tích từ vựng (lexical analysis) là bước đầu tiên của trình dịch
  - Còn gọi là scanning hoặc lexing, bộ phân tích từ vựng là scanner hoặc lexer
- Khối phân tích từ vựng (PTTV):
  - Nhận dữ liệu đầu vào là mã nguồn cần dịch
  - Loại bỏ các đoạn mã không cần thiết
  - Chia đoạn mã còn lại thành dãy các từ tố (token)
  - Chuyển kết quả cho khối phân tích cú pháp (PTCP)
- Tương tác giữa PTTV và PTCP như thế nào?

# 1. Vai trò của phân tích từ vựng

- Có nhiều quan điểm về sự tương tác giữa bộ PTTV và bộ phân tích cú pháp.
  - Thiết kế cổ điển: coi PTTV như một tiến trình phụ thuộc vào bộ phân tích cú pháp, quá trình phân tích cú pháp điều khiển việc phân tích từ vựng.
  - Thiết kế hiện đại: tách PTTV thành một module độc lập, kết quả đầu ra của PTTV được tiêu chuẩn hóa để có thể được ghi ra file hoặc sử dụng bởi các mục đích khác.
- Chú ý: việc chọn cách thiết kế là do mục tiêu xây dựng chương trình, không có nghĩa là thiết kế hiện đại thì tốt hơn thiết kế cổ điển.

# 1. Vai trò của phân tích từ vựng

- Trong thiết kế cổ điển, PTTV đóng vai trò như bộ cung cấp dữ liệu cho bộ phân tích cú pháp.
  - Bộ phân tích cú pháp yêu cầu PTTV lấy từ tố tiếp theo.
  - Bộ PTTV đọc chương trình nguồn từ đầu hoặc từ vị trí đang phân tích trong lần gọi trước, tách lấy từ tố tiếp theo trả lại cho bộ phân tích cú pháp.
  - Quá trình lặp lại cho đến khi hết mã nguồn hoặc gặp lỗi.



# 1. Vai trò của phân tích từ vựng

PHENIKAA UNIVERSITY
Realizing your potential

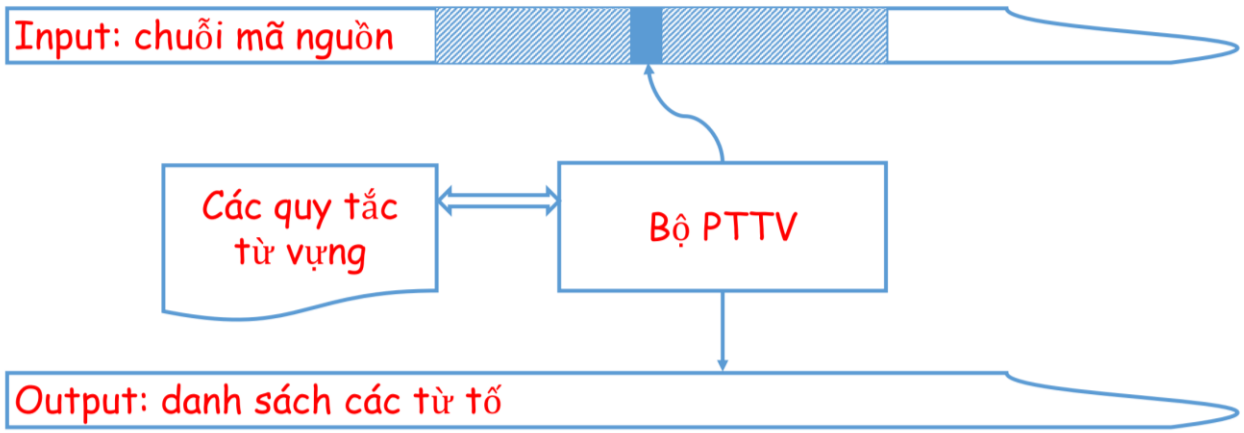
- Trong các thiết kế mới hơn, bộ PTTV có xu hướng đứng tách ra độc lập, việc này có nhiều lợi ích:
  - Thiết kế theo hướng module hóa, đơn giản hơn.
  - Tăng hiệu quả hoạt động của bộ PTTV, chẳng hạn như PTTV có thể độc lập xử lý các macro, xử lý khoảng trắng, ghi chú,...
  - Tối ưu hoạt động của trình dịch, bộ PTTV sau khi hoạt động có thể giải phóng các tài nguyên mà nó sử dụng thay vì giữ lại cùng lúc với bộ phân tích cú pháp.
  - Xử lý được ngay lập tức một số lỗi cơ bản về từ vựng mà không cần phân tích cú pháp.

# 2. Nhiệm vụ của phân tích từ vựng

- PTTV đóng vai trò như một bộ chuẩn hóa dữ liệu đầu vào, ngoài ra nó cũng giúp hạn chế các lỗi cơ bản (viết sai luật, sai từ khóa, sai cấu trúc,...)
- Các nhiệm vụ chính (nhất thiết phải có để đảm bảo hoạt động của chương trình dịch):
  - Đọc chương trình nguồn, loại bỏ các kí hiệu vô ích (khoảng trắng, dấu tab, xuống dòng, ghi chú,...).
  - Phát hiện một số lỗi cơ bản về từ vựng.
  - Xác định nội dung của từ vựng.
  - Xác định từ loại của từ vựng đó.
  - Đưa ra một số thông tin thuộc tính của từ vựng.

# 3. Các bước để phân tích từ vựng

- Đầu vào của PTTV: Trong trường hợp tổng quát nhất, đầu vào của bộ PTTV là mã nguồn cần phân tích, không có bất kì ràng buộc nào.
- Đầu ra của bộ PTTV: phụ thuộc vào các đặc điểm của ngôn ngữ nguồn và bộ phân tích cú pháp.
  - Trong hầu hết các tình huống, bộ PTTV thường trả về kết quả ở dạng sau:
  - Danh sách các từ vựng ứng theo mã nguồn (thường là một danh sách liên kết, chẳng hạn – List<Word>)
    - Với mỗi từ vựng, thông tin bao gồm:
    - Từ loại của từ vựng.
    - Giá trị chính xác của từ vựng.
    - Giá trị mã hóa của từ vựng.
    - Vị trí của từ vựng trong mã nguồn.



# 3. Các bước để phân tích từ vựng

PHENIKAA UNIVERSITY
Realizing your potential

- Các bước thực hiện:
  1) Xoá bỏ các kí tự không có nghĩa: chú thích, dòng trống, các ký tự xuống dòng, dấu tab, các khoảng trắng không cần thiết.
  2) Nhận dạng các ký hiệu: các kí tự liền nhau tạo thành một kí hiệu. Các dạng kí hiệu gọi là các từ tố.
  3) Số hoá ký hiệu.

# 3. Các bước để phân tích từ vựng

PHENIKAA UNIVERSITY
Realizing your potential

- Ví dụ:
  - *Đầu vào:*
    *position := initial + rate * 60;*
  - Đầu ra:

<tên, con trỏ đến position trên bảng ký hiệu>
<phép_gán , >
<tên, con trỏ đến initial trên bảng ký hiệu>
<toán_tử_cộng , >
<tên, con trỏ đến rate trên bảng ký hiệu>
<toán_tử_nhân, >
<số, giá trị số nguyên 60>
<chấm_phẩy, >

# 3. Các bước để phân tích từ vựng

PHENIKAA UNIVERSITY
Realizing your potential

❑Ví dụ:

| Từ tố (token) | Từ vị | Mẫu (luật mô tả) |
|---|---|---|
| **const** | const | const |
| **if** | if | if |
| **quan hệ** | <, <=, =, <>, >, >= | <, <=, =, <>, >, >= |
| **tên** | pi, count, i, d2 | Chữ cái theo sau là các chữ cái hoặc số |
| **số** | 3.14, 0, 6.02E23 | Bất cứ một hằng số nào đó |
| **xâu** | "Xin chao cac ban" | Bất cữ chữ nào đặt trong dấu ", trừ dấu " |

# 3. Các bước để phân tích từ vựng

PHENIKAA UNIVERSITY
Realizing your potential

- Thuộc tính của các từ tố.
  - Một từ tố có thể ứng với một tập các từ vị khác nhau, phải thêm một số thông tin nữa để khi cần có thể biết được cụ thể đó là từ vị nào.

# 3. Các bước để phân tích từ vựng

PHENIKAA UNIVERSITY
Realizing your potential

Ví dụ

- Cho biểu thức

  **position := initial + rate * 60;**

- Dãy từ tố nhận được bao gồm:

  <tên, con trỏ đến position trên bảng kí hiệu>
  <phép_gán, >
  <tên, con trỏ đến initial trên bảng kí hiệu>
  <toán_tử_cộng, >
  <tên, con trỏ đến rate trên bảng kí hiệu>
  <toán_tử_nhân, >
  <số nguyên, giá trị số nguyên 60>
  <chấm phẩy, >

# 4. Xác định từ tố

- Biển diễn từ tố:
  - Các từ tố khác nhau có các luật mô tả (mẫu) khác nhau. Các luật mô tả này là cơ sở để nhận dạng được từ tố.
  - Cách biểu diễn các luật này đơn giản và thông dụng nhất là bằng lời. Nhược: nhập nhằng.
  - Biểu diễn tốt nhất: dùng biểu thức chính quy và ôtômát hữu hạn - lớp ngôn ngữ chính quy.

# 4. Xác định từ tố: Biểu thức chính quy

- Ví dụ biểu diễn các khái niệm về chữ cái, chữ số, tên, phép quan hệ trong Pascal như sau:

  chữ cái → A | B | ... | Z | a | b | ... | z

  chữ số → 0 | 1 | ... | 9

  tên → chữcái ( chữcái | chữsố )*

  quan hệ → < | <= | = | <> | > | >=

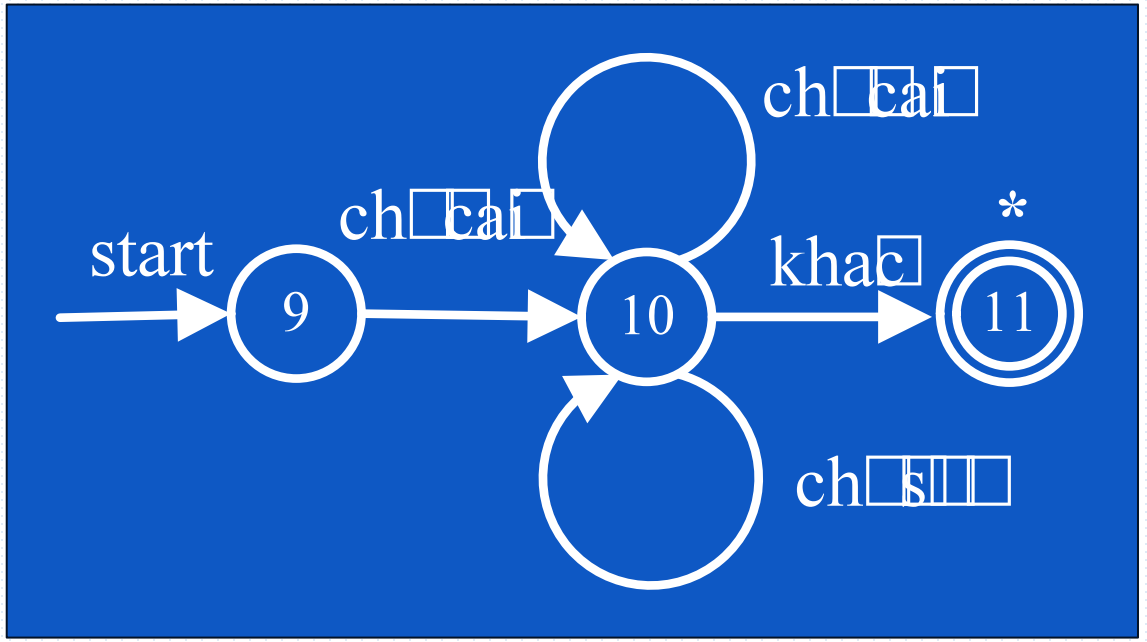
- Các qui ước bổ sung:

  \+ lặp lại một hoặc nhiều lần

  ? lặp lại không hoặc một lần

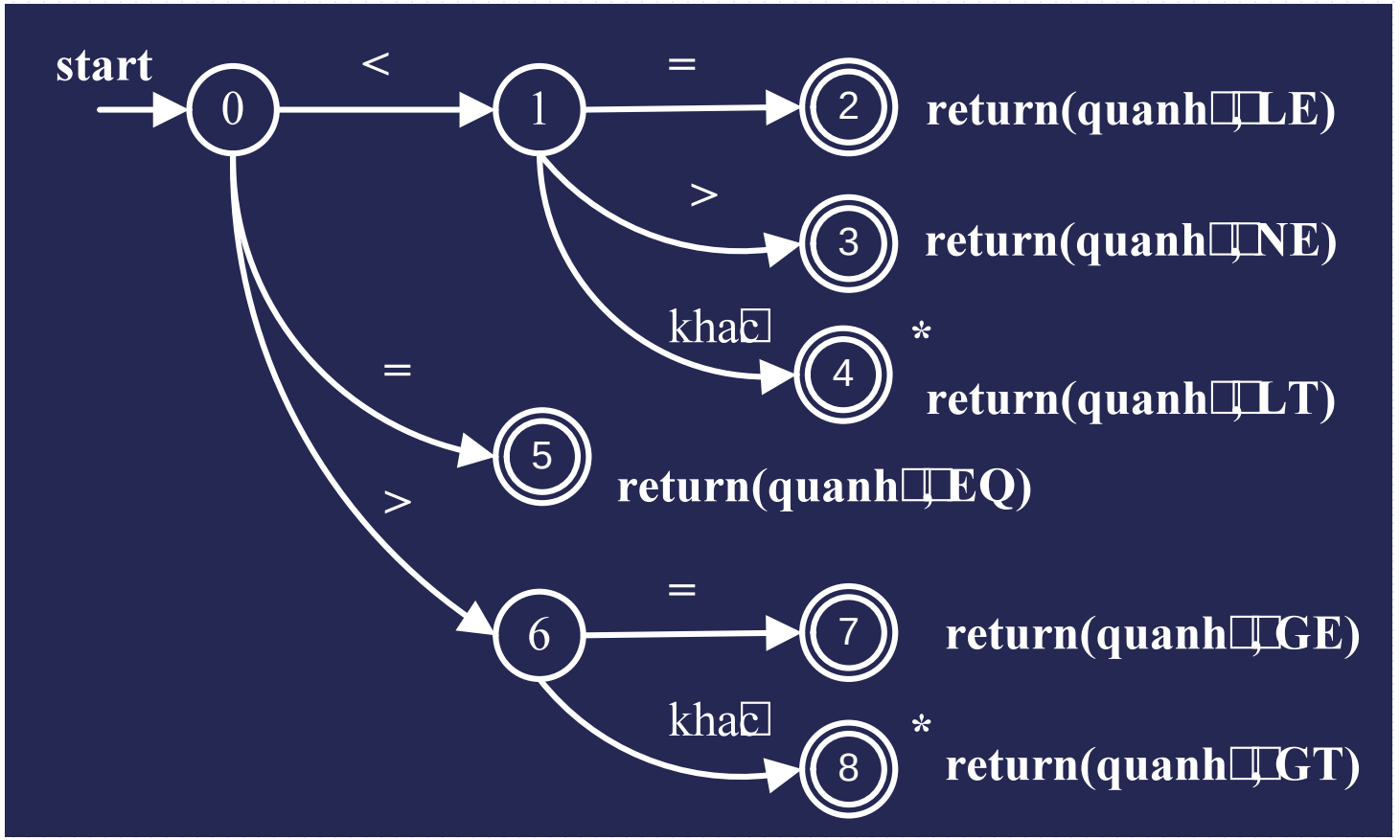
# Đồ thị chuyển

- Đồ thị chuyển mô tả một ô tô mát đoán nhận xâu đầu vào là một tên.
- Mỗi một ngôn ngữ khác nhau có cách sử dụng tên/toán tử/quan hệ khác nhau → đồ thị chuyển khác nhau.
- Ví dụ: Biểu diễn tên
  - Ký tự * đồ thị đã xử lý quá một ký hiệu của phần khác.



# Đồ thị chuyển

- Ví dụ: Đồ thị biểu diễn quan hệ



# Viết chương trình cho đồ thị chuyển

- Tập hợp tất cả các mẫu của từ tố
  - Tên, từ khóa, quan hệ, vv.
- Lập bộ phân tích từ vựng bằng phương pháp diễn giải đồ thị chuyển.
- Lập bộ phân tích từ vựng điều khiển bằng bảng (mô phỏng ôtômát hữu hạn đơn định).

# Tập hợp tất cả các mẫu của từ tố

PHENIKAA UNIVERSITY
Realizing your potential

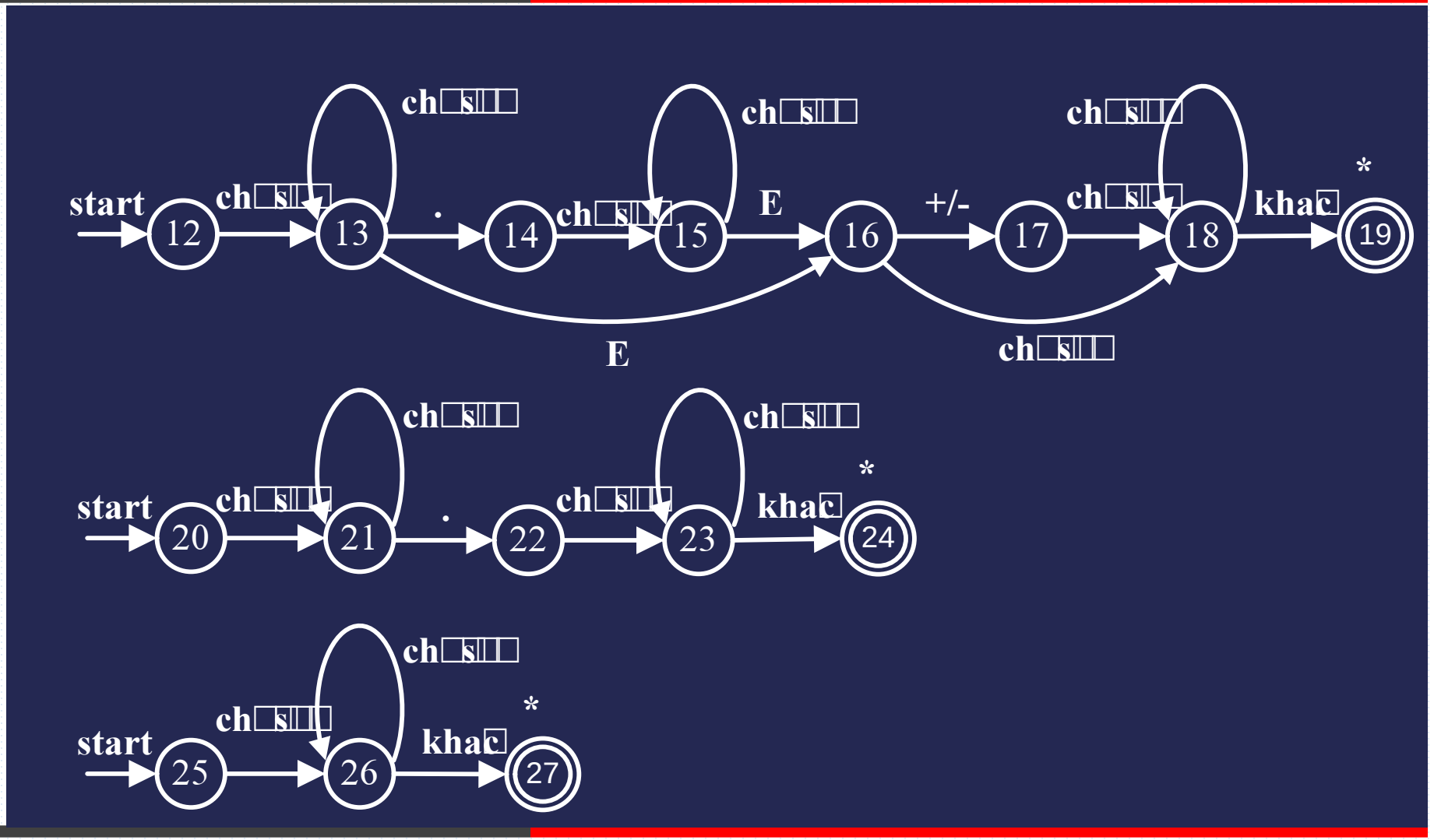
- Ví dụ: Biểu diễn số

  sốthựcmũ → chữsố+ (. chữsố+ ) ? ( E ( + | - ) ? chữsố+) ?

  sốthực → chữsố+ . chữsố+

  sốnguyên → chữsố+

# Tập hợp tất cả các mẫu của từ tố



# Lập bộ phân tích từ vựng bằng phương pháp diễn giải đồ thị chuyển

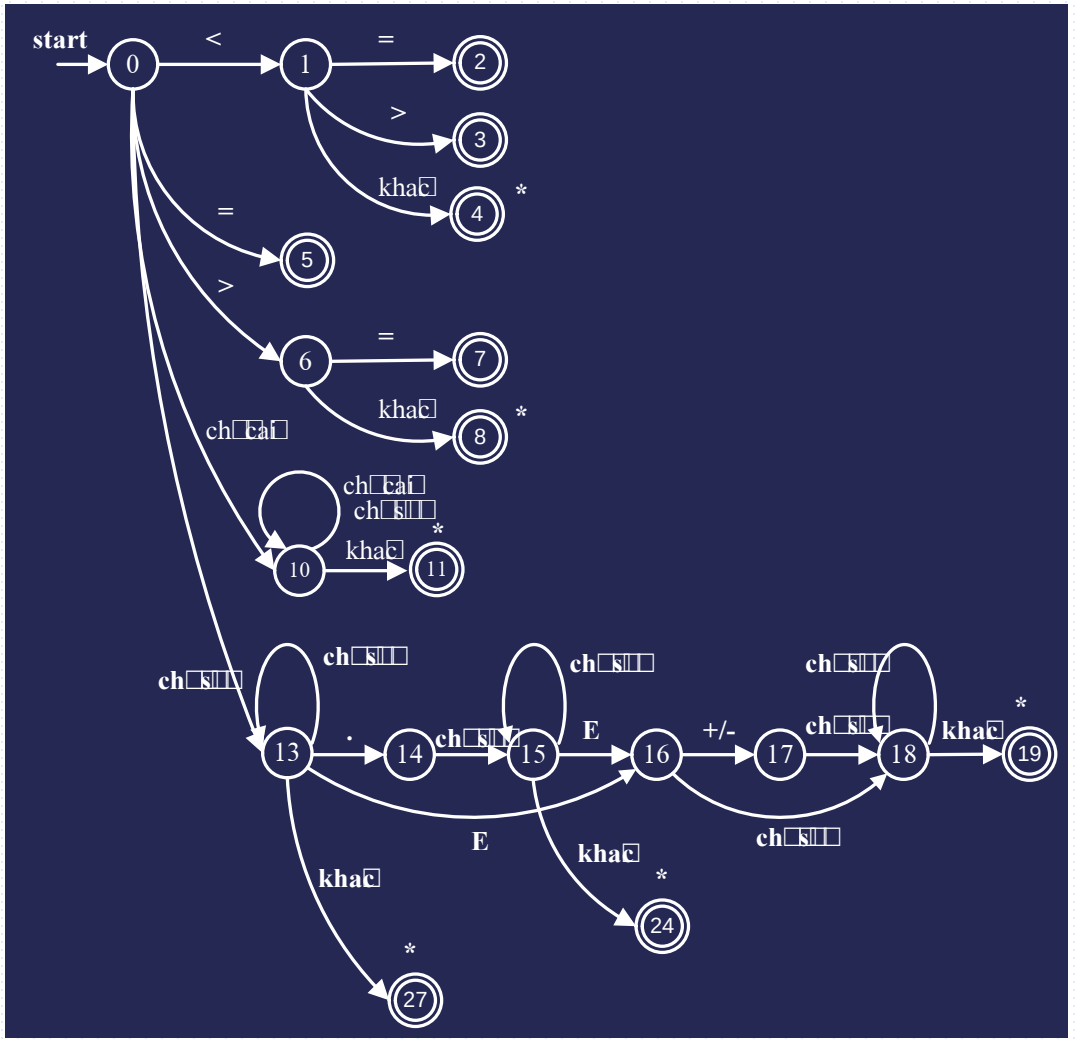
- Lần theo sơ đồ, dùng các câu lệnh lựa chọn if hoặc switch
- Ưu điểm: dễ hiểu, dễ viết.
- Nhược điểm: gắn kết cấu đồ thị chuyển vào trong chương trình. Khi thay đổi đồ thị thì phải viết lại chương trình nên khó bảo trì.

# Lập bộ phân tích từ vựng điều khiển bằng bảng

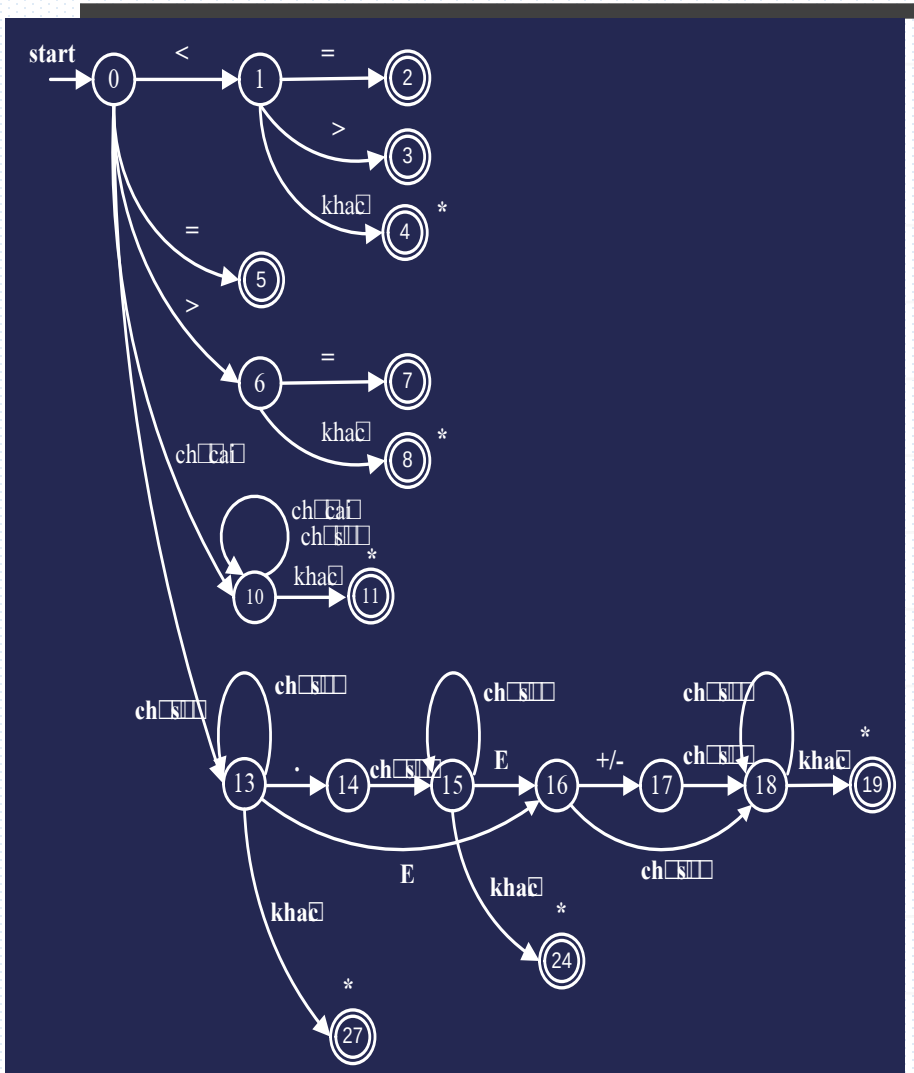
- Kết hợp các đồ thị chuyển thành một đồ thị chuyển duy nhất

# Lập bộ phân tích từ vựng điều khiển bằng bảng

- Kết hợp các đồ thị chuyển thành một đồ thị chuyển duy nhất
- Đồ thị chuyển
  - 24 trạng thái
  - Các kí tự vào từ chương trình nguồn được chia thành các loại: chữcái, chữsố, E, ., <, =, >,+/- (tức là có 8 loại khác nhau).
  - Lập một bảng 10 x 8



# Lập bộ phân tích từ vựng điều khiển bằng bảng




| Trạng thái | Loại kí tự vào | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | < | = | > | chữ cái | chữ số | E | . | +/- |
| 0 | 1 | 5 | 6 | 10 | 13 | | | |
| 1 | 4 | 2 | 3 | 4 | 4 | 4 | 4 | 4 |
| 6 | 8 | 7 | 8 | 8 | 8 | 8 | 8 | 8 |
| 10 | 11 | 11 | 11 | 10 | 10 | 11 | 11 | 11 |
| 13 | 27 | 27 | 27 | 27 | 13 | 16 | 14 | 27 |
| 14 | | | | | 15 | | | |
| 15 | 24 | 24 | 24 | 24 | 15 | 16 | 24 | 24 |
| 16 | | | | | 18 | | | 17 |
| 17 | | | | | 18 | | | |
| 18 | 19 | 19 | 19 | 19 | 18 | 19 | 19 | 19 |

# Viết chương trình

- Dựa vào bảng chuyển để chuyển trạng thái
- Ưu điểm: tách dữ liệu độc lập với chương trình, do đó rất dễ biến đổi mà không cần phải sửa lại chương trình.
- Nhược điểm: là khó hiểu hơn và khó lập bảng

# Cách xử lý lỗi

- Đơn giản nhất: hệ thống sẽ ngừng hoạt động và báo lỗi cho người sử dụng.
- Tốt hơn và hiệu quả hơn: ghi ra các lỗi này và cố gắng bỏ qua chúng để tiếp tục làm việc, nhằm phát hiện đồng thời thêm nhiều lỗi khác.
- Các cách khắc phục có thể có:
  - Xoá hoặc nhẩy qua các kí tự mà bộ phân tích không tìm được từ tố;
  - Thêm một kí tự bị thiếu;
  - Thay một kí tự sai bằng một kí tự đúng;
  - Tráo hai kí tự đứng cạnh nhau.

# 5. Các bước để xây dựng một bộ PTTV

- Sưu tầm tất cả các luật từ vựng, các luật này thường được mô tả bằng lời.
- Vẽ đồ thị chuyển cho từng luật một. Trước đó có thể mô tả chúng bằng các biểu thức chính quy để tiện theo dõi và chỉnh sửa, và làm dễ cho việc dựng đồ thị.
- Kết hợp các luật này thành một đồ thị chuyển duy nhất.
- Chuyển đồ thị này thành bảng.
- Thêm phần chương trình ở trên để thành bộ phân tích hoạt động được.
- Thêm phần báo lỗi để thành bộ phân tích từ vựng hoàn chỉnh

# Bài tập

1. Mô tả các ngôn ngữ chỉ định bởi các biểu thức chính quy sau:

a. $0(0|1)^*0$

b. $(\epsilon|0)1^*)^*$

2. Viết biểu thức chính quy cho: tên, số nguyên, số thực, char, string. Sau đó kết hợp chúng thành một đồ thị chuyển duy nhất

3. Dựng đồ thị chuyển cho các mô tả dưới đây

- Tất cả các xâu chữ cái có 6 nguyên âm a, e, i, o, u, y theo thứ tự.
  + Ví dụ: "abeihgtry".
- Tất cả các xâu số không có một số nào bị lặp.
- Tất cả các xâu số có ít nhất một số nào đó bị lặp.
- Tất cả các xâu chỉ bao gồm các chữ số 0 và 1, nhưng không chứa xâu con 011.
- Tất cả các xâu chỉ bao gồm các chữ số 0 và 1, nhưng không chứa liên tục các xâu con 011.