**Semester project report**

**Theory of automata and formal languages**

**Project Title:**

Natural Language Identifier

**Abstract:**

This project presents a simulation of a **Natural Language Identifier** that uses concepts from automata theory, particularly **Deterministic Finite Automata (DFA)**.  
Words from multiple languages are matched using finite sets, and the system determines which language the input text most likely belongs to.  
The project implements DFA simulation through **Python** and provides a user-friendly  **GUI** for easy interaction.

**Purpose:**

* To apply the theory of finite automata in real-world applications.
* To demonstrate how a DFA can be used to recognize patterns and classify natural languages based on keywords.
* To develop an educational tool that bridges theoretical computer science and practical programming.

**Applications:**

* Detecting language before translation.
* Language-based text categorization.
* Preprocessing for NLP tasks like chatbots, sentiment analysis.

**Project Description:**

The Natural Language Identifier uses a DFA-like approach to recognize predefined **sets of words** (patterns) for each language.  
When a user inputs a text, it is cleaned (lowercased and punctuation removed), tokenized into words, and compared against these sets.  
The language with the highest number of matched words is predicted as the input language.  
A simple yet attractive **GUI** built with Tkinter allows users to interact with the system.

**RE (Regular Expressions):**

For **each language**, the Regular Expression (RE) is simply a union (+) of specific words:

**English RE:**  
(the + is + and + hello + how + are + you + good + morning + night)

**Spanish RE:**  
(el + es + y + hola + como + estas + buenos + dias)

**French RE:**  
(le + est + et + bonjour + comment + ça + va + merci)

**German RE:**  
(der + ist + und + hallo + wie + gehts + danke)

**Italian RE:**  
(il + è + e + ciao + come + stai + grazie + buongiorno)

**Hindi RE:**  
(नमस्ते + कैसे + हो + आप + शुभ + सुप्रभात)

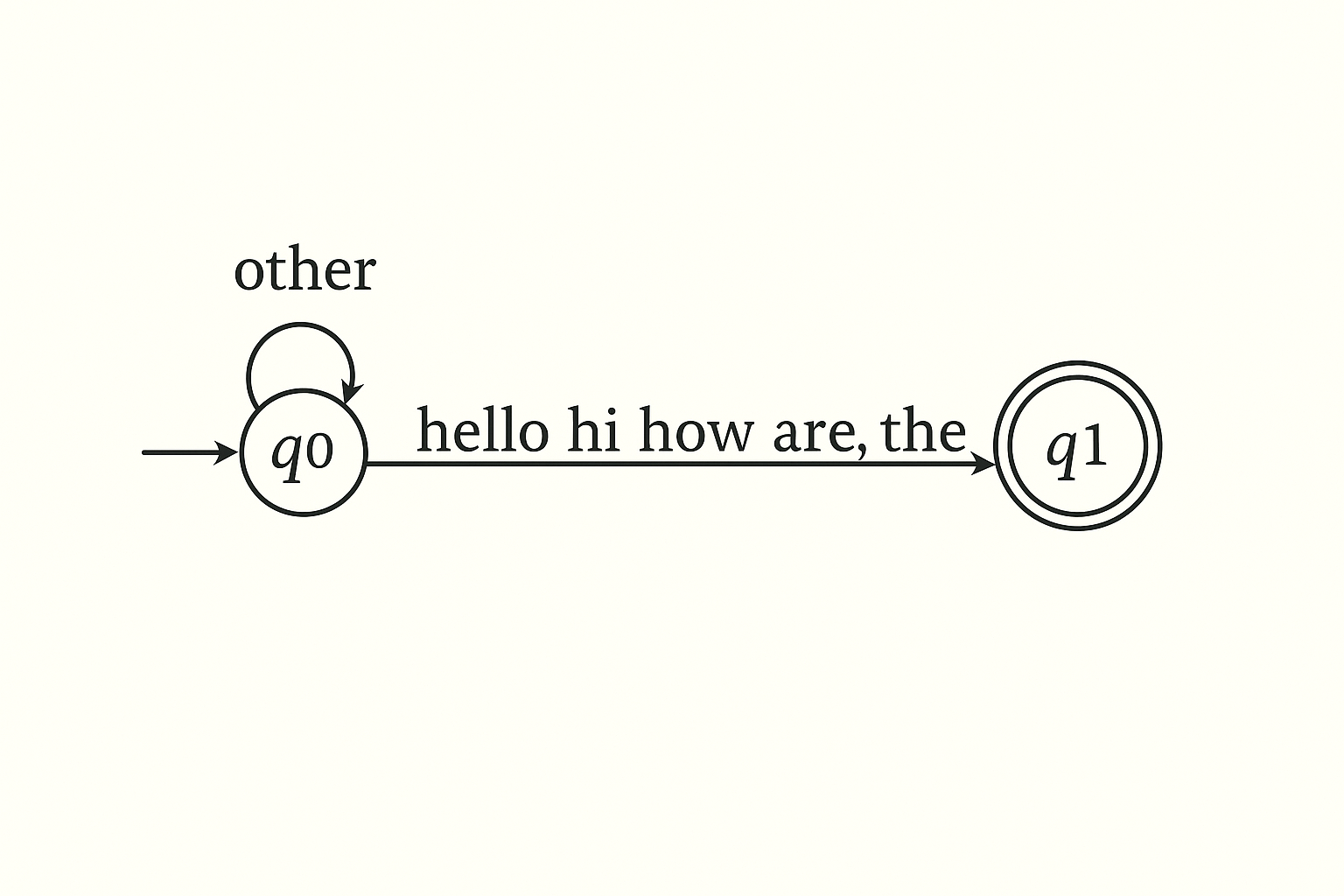
**Urdu RE:**  
( ہے + وہ + یہ +ہیلو + آپ + کیسے + ہیں + صبح + اچھی + شب + بخیر + کا + کی + کے)

**Arabic RE:**  
(مرحبا + كيف + حالك + السلام + عليكم + صباح + الخير)

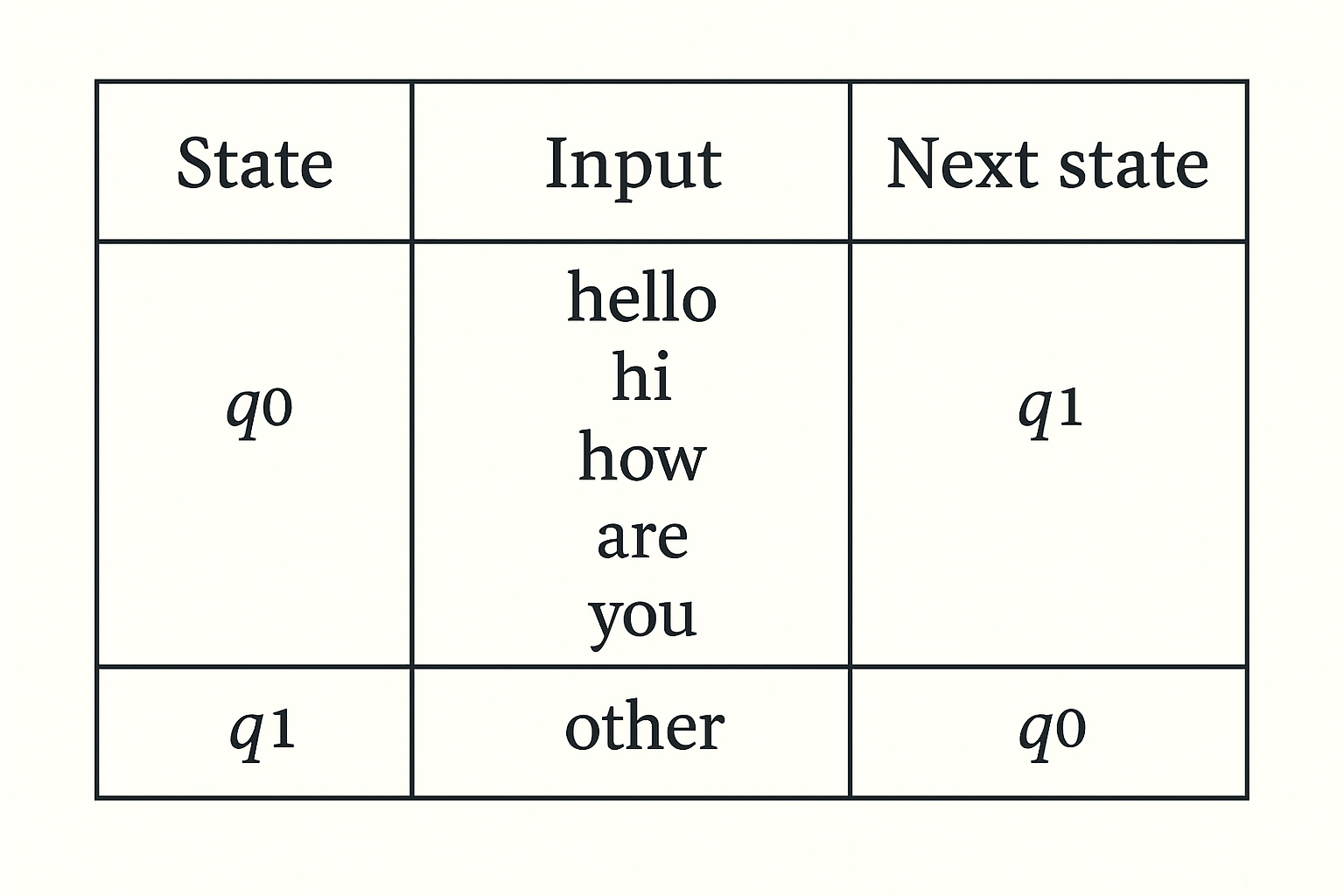
**FA (Finite Automaton):**

A **DFA** for each language can be like this:

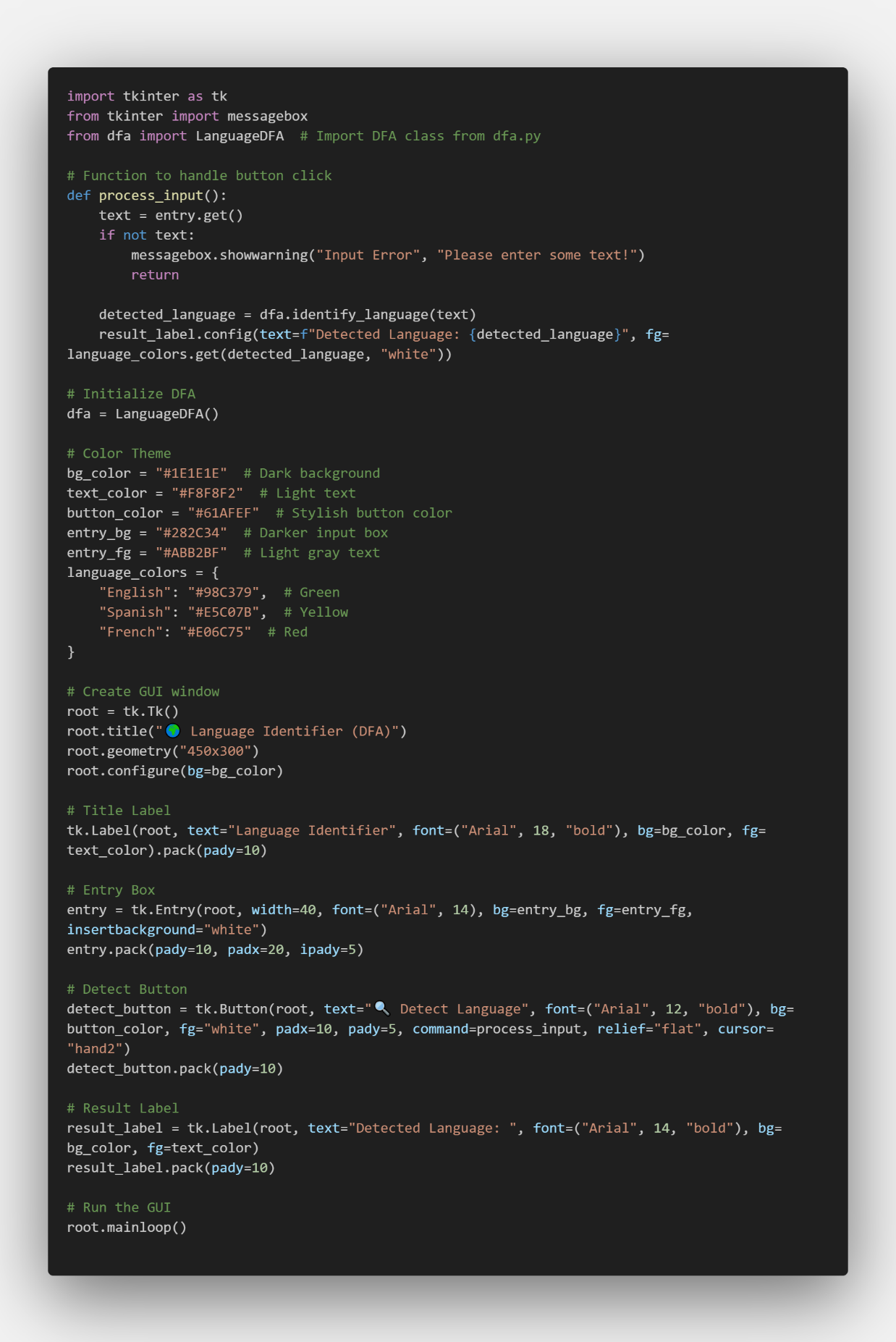
* Start at **q0**.
* If the first word matches any keyword of a language, move to **q1** (accepting state for that language).
* If not matched, remain in **q0**.
* Each successful keyword match strengthens the prediction for that language.



**Transition Graph (TG):**



**Code:**

****

****