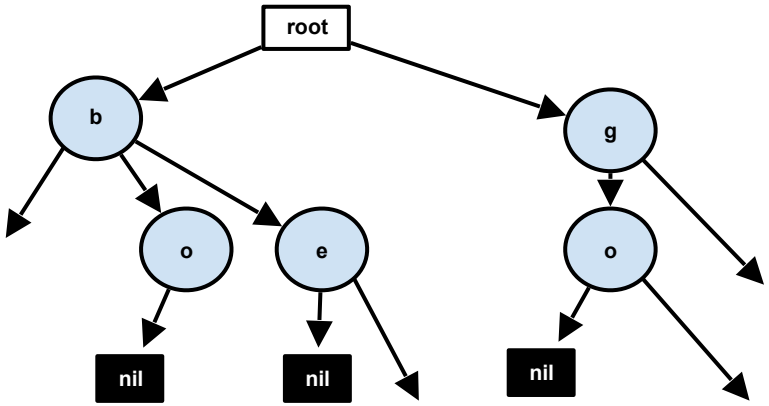
Prefix Trees (Tries) for Tamil Language Processing

Elango Cheran
Tamil Internet Conference
Aug. 2017
Toronto, Canada

இளங்கோ சேரன்
தமிழ் இணைய மாநாடு
ஆ. 2017
தொராண்டோ, கனடா

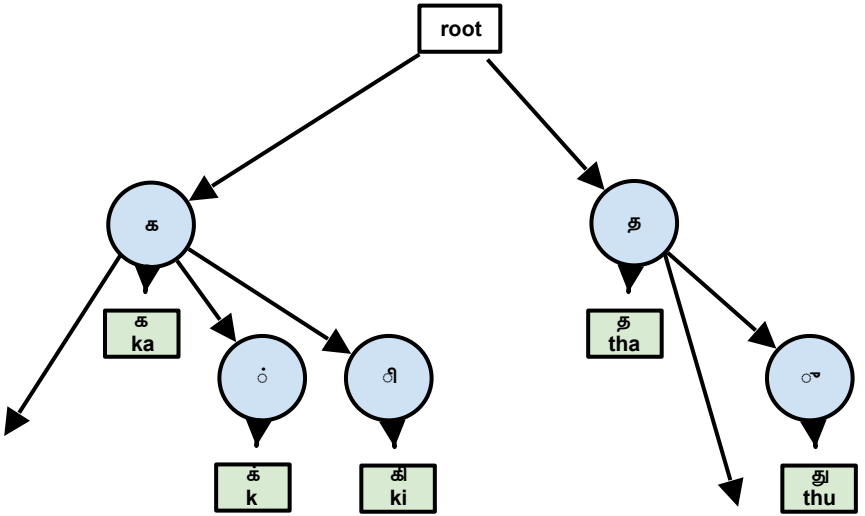
What is a prefix tree?

- A type of tree data structure



prefix tree என்றால் எனன்?

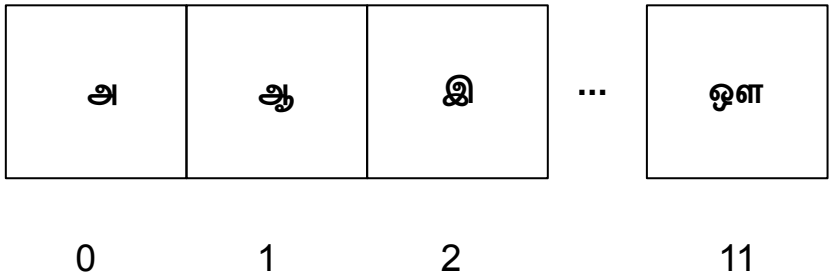
- ஒரு வகையான data structure



Data Structures - Vector/ Array

- Sequence of elements
- Indexed - can get by position number
- Fast amortized append / delete

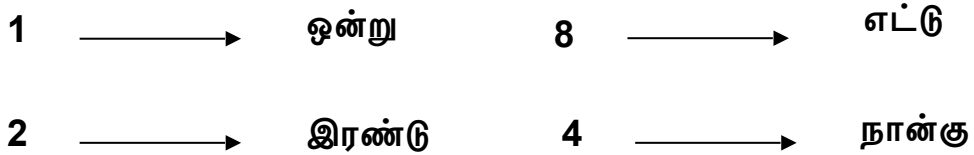
- தனிமங்கள் வரிசையில் இருத்தல்
- Indexed - வரிசை எண் வைத்து எடுக்கலாம்
- விரைவாகச் சேர்த்தல், கழித்தல்



Data Structures - Map

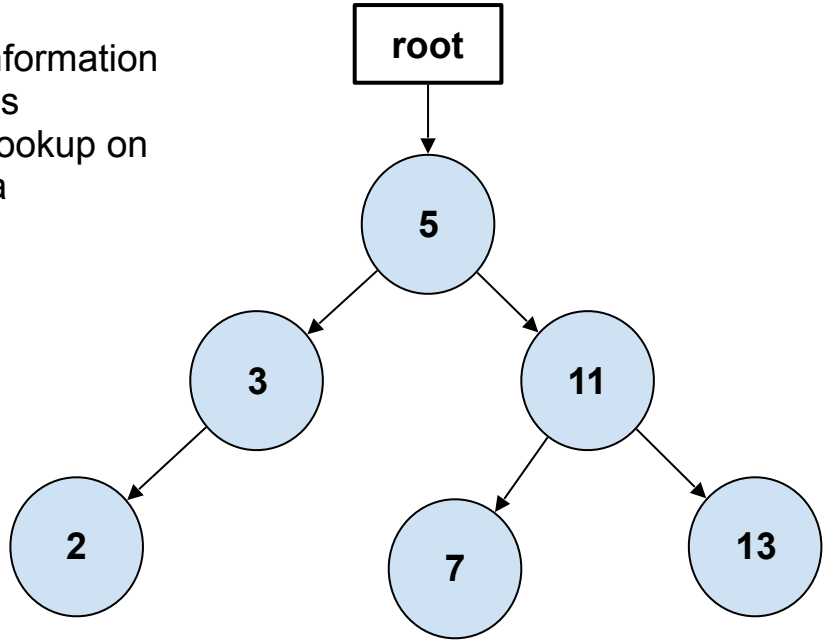
- Associates keys to values
- No ordering
- Fast lookup / put / removal

- Key-களை value-களோடு தொடர்புபடுத்துதல்
- வரிசை இல்லை
- விரைவான கண்டுபிடித்தல், போடுதல், கழித்தல்



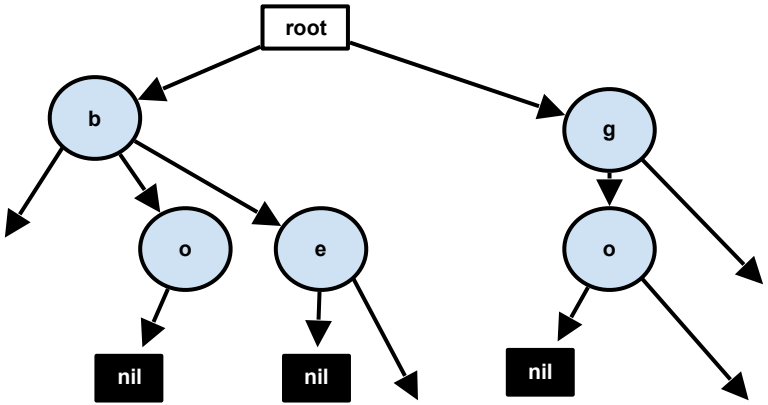
Data Structures - Tree

- Graph (nodes, edges)
- Root node
- Store a value in a new node
- Examples
  - hierarchical information
  - decision states
  - sorting / fast lookup on changing data



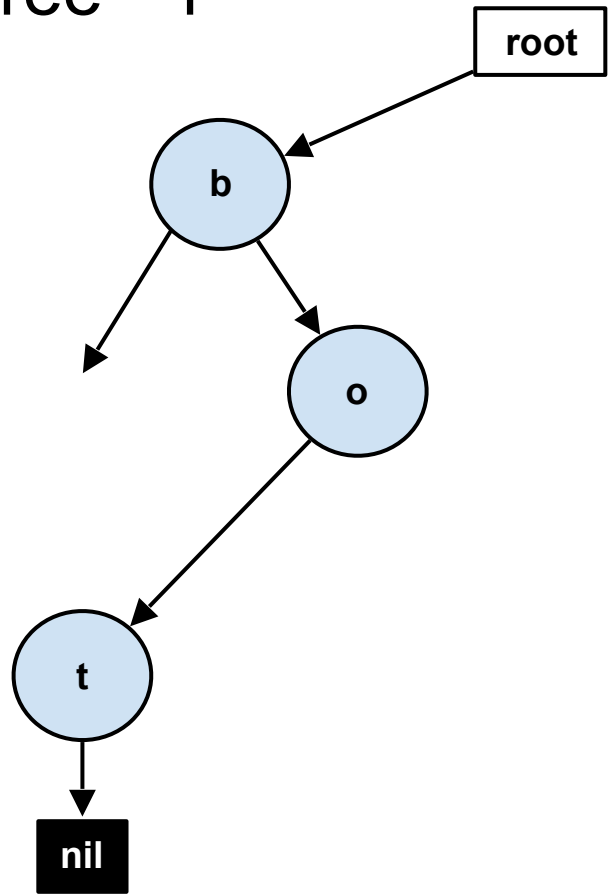
Data Structures - Prefix Tree

- Graph (nodes, edges)
- Root node
- Store a value (sequence) as a path from the root
- Very basic idea for NLP work
- Examples
  - Storing a vocabulary of words from a natural language



Building a prefix tree - 1

- Adding [b, o, t]

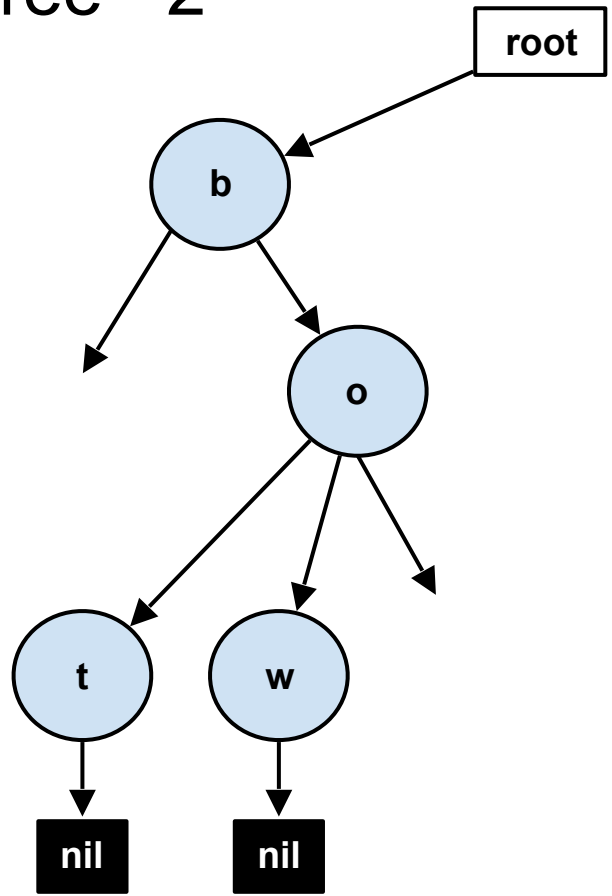


prefix tree அமைத்தல் - 1

- [b, o, t] சேர்க்கிறது

Building a prefix tree - 2

- Adding [b, o, w]



prefix tree
அமைத்தல் - 2

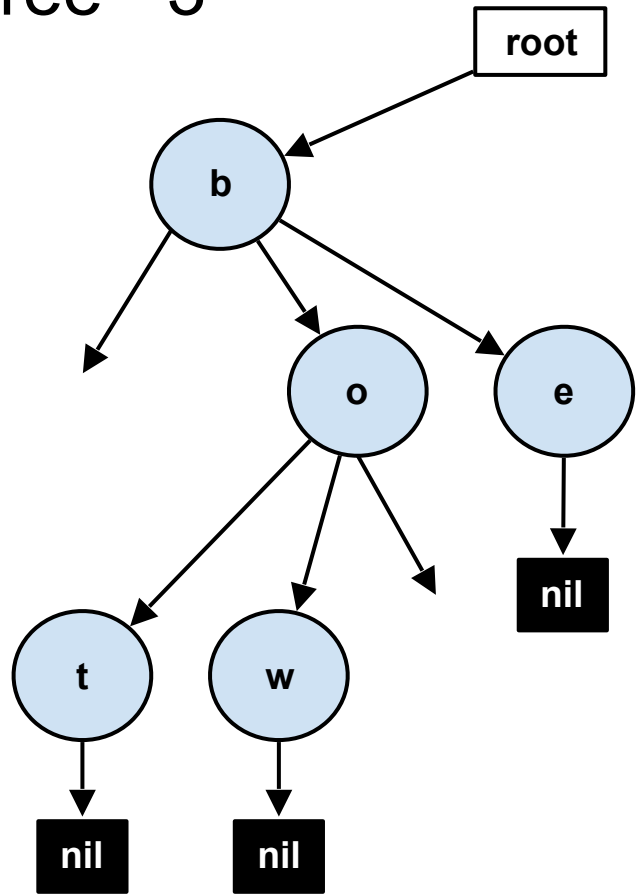
- [b, o, ந்] சேர்க்கிறது

Building a prefix tree - 3

- Adding [b, e]

prefix tree அமைத்தல் - 3

- [b, e] சேர்க்கிறது

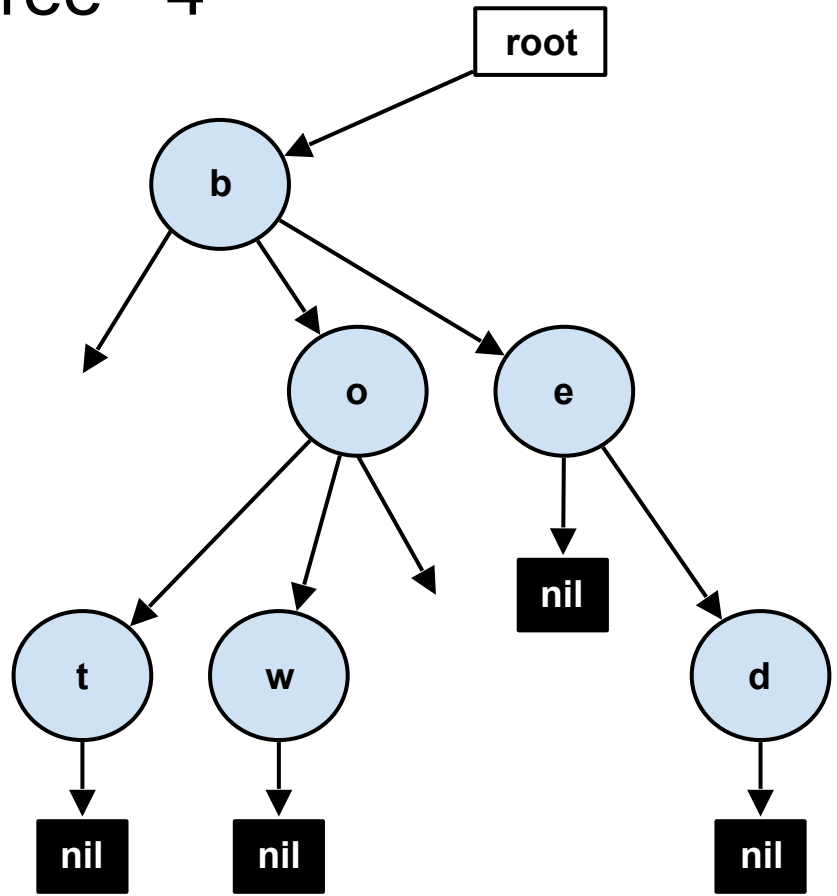


Building a prefix tree - 4

- Adding [b, e, d]

prefix tree அமைத்தல் - 4

- [b, e, d] சேர்க்கிறது

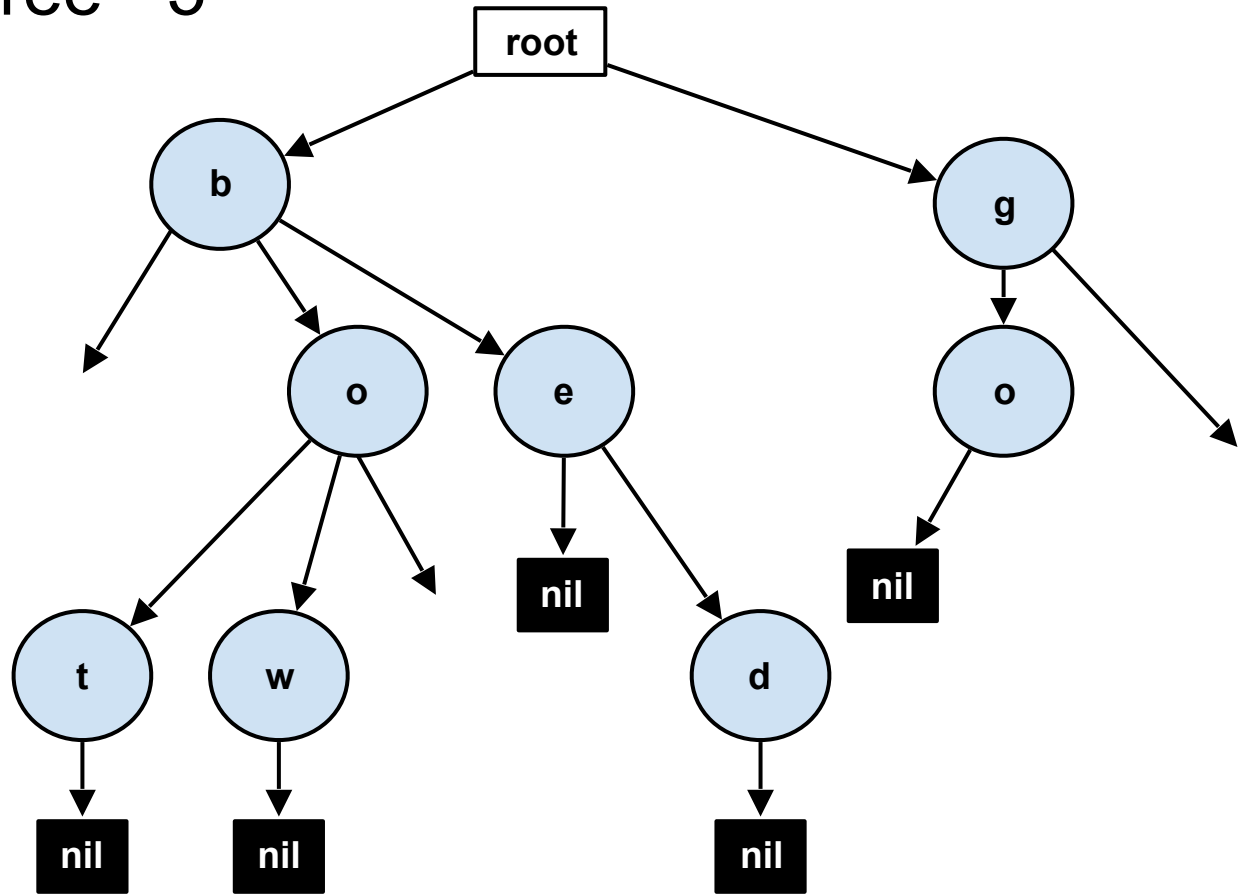


Building a prefix tree - 5

- Adding [g, o]

prefix tree அமைத்தல் - 5

- [g, o] சேர்க்கிறது

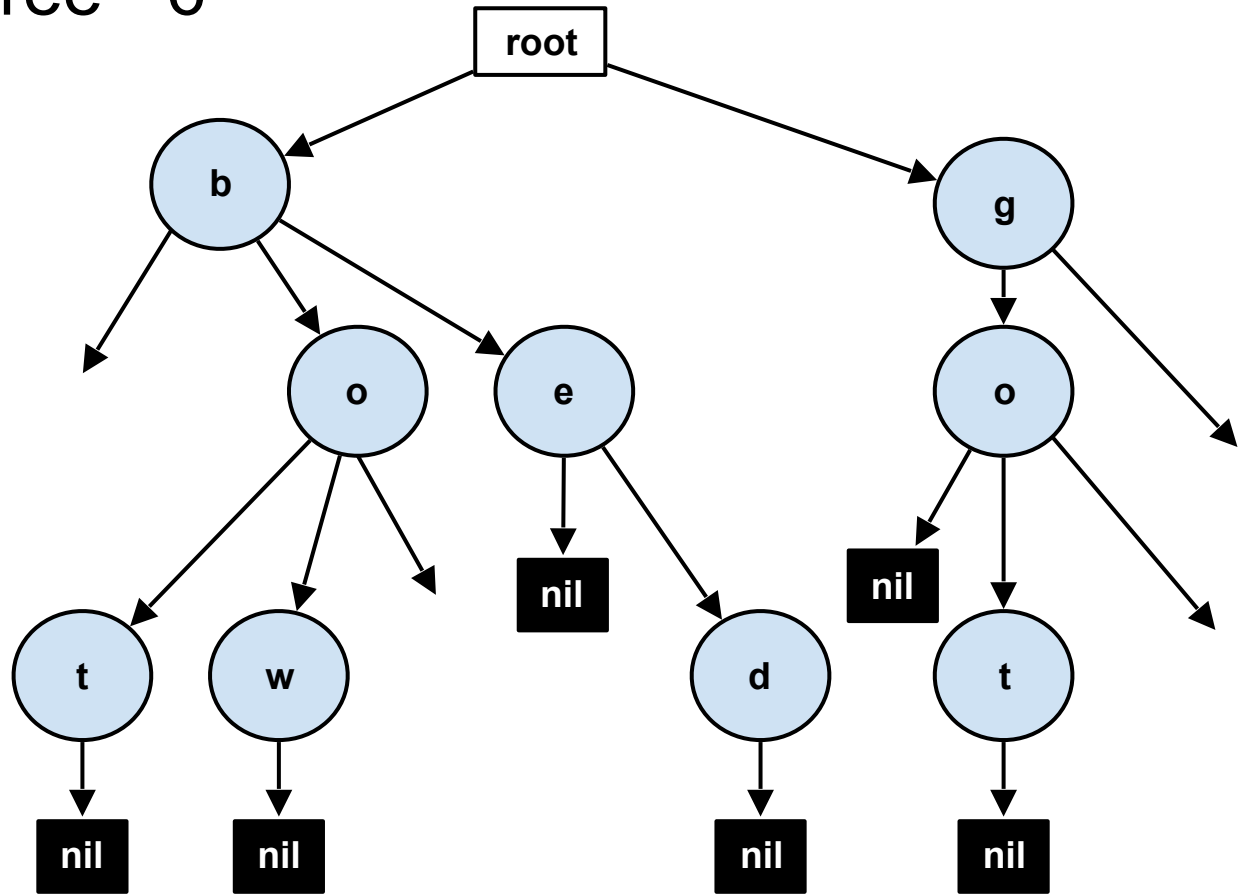


Building a prefix tree - 6

- Adding [g, o, t]

prefix tree அமைத்தல் - 6

- [g, o, t] சேர்க்கிறது



Unicode for Tamil has solvable challenges

தமிழ் ஒற்றைக்குறியீட்டில் தீர்க்கக்கூடிய சவால்கள்

Text	Logical letters	Number of logical letters	Text Unicode codepoints	Number of Text Unicode codepoints
go	g, o	2	g, o	2
got	g, o, t	3	g, o, t	3
bot	b, o, t	3	b, o, t	3
bow	b, o, w	3	b, o, w	3
தணி	த, ணி	2	த, ண, ி (0BBF)	3
தணிகை	த, ணி, கை	3	த, ண, ி (0BBF), க, ை (0BC8)	5
வருகை	வ, ரு, கை	3	வ, ர, ு (0BC1), க, ை (0BC8)	5
வருக	வ, ரு, க	3	வ, ர, ு (0BC1), க	4

Prefix Trees for Tamil

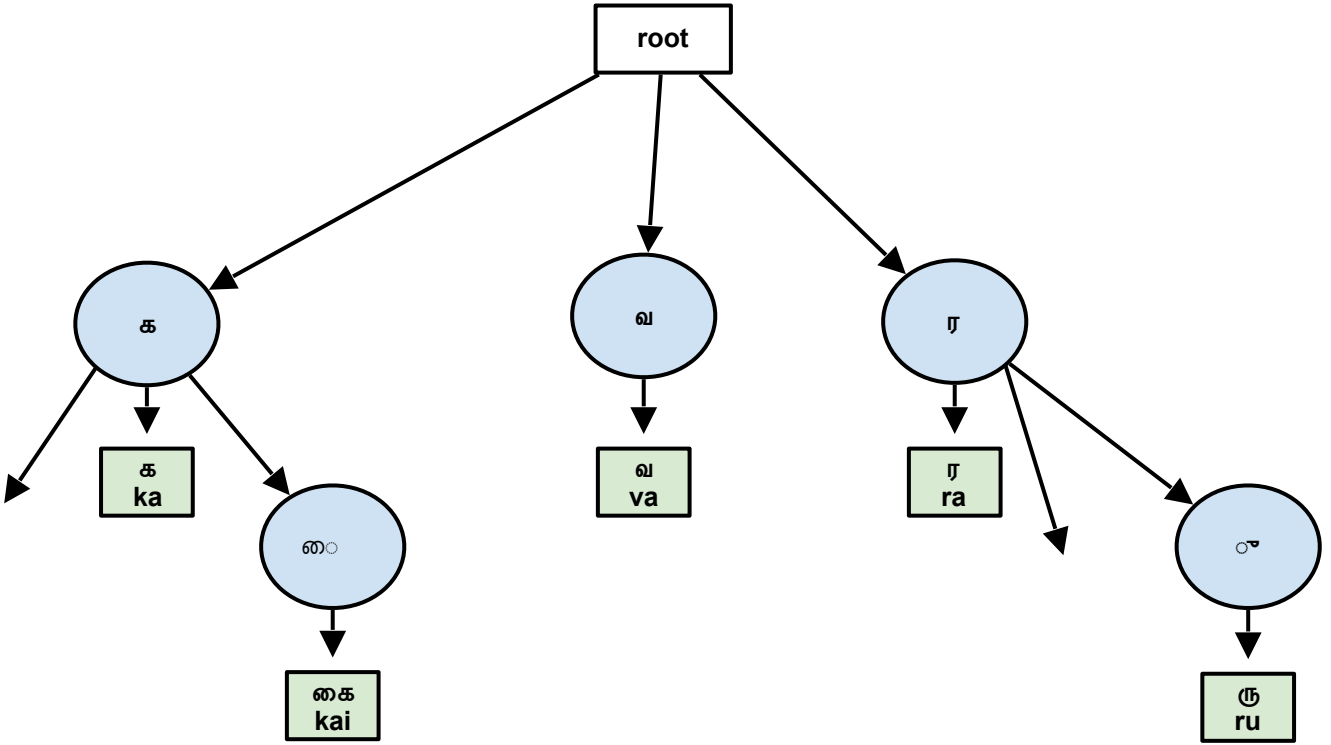
- Directly useful for:
  - Parsing text into letters
  - Parsing text into phonemes
  - Converting phoneme sequences back to letters
  - Transliteration
- Also enables:
  - Functions performing grammatical changes

- நேரடியாகப் பயன்படும்:
  - உரையை எழுத்துகளாகப் பிரித்தல்
  - உரையை ஒலியன்களாகப் பிரித்தல்
  - ஒலியன்களை மீண்டும் எழுத்துகளாக மாற்றுதல்
  - எழுத்துரு மாற்றம் / எழுதும் மொழி மாற்றம்
- இதன் மூலம் இவை விளைவாக வரும்:
  - இலக்கண மாற்றங்களைச் செயல்படுத்தும் செயல்கூறுகள்

Prefix Tree Parsing Tamil Letters

Input:
வ, ர, ◌ை, க, ை◌

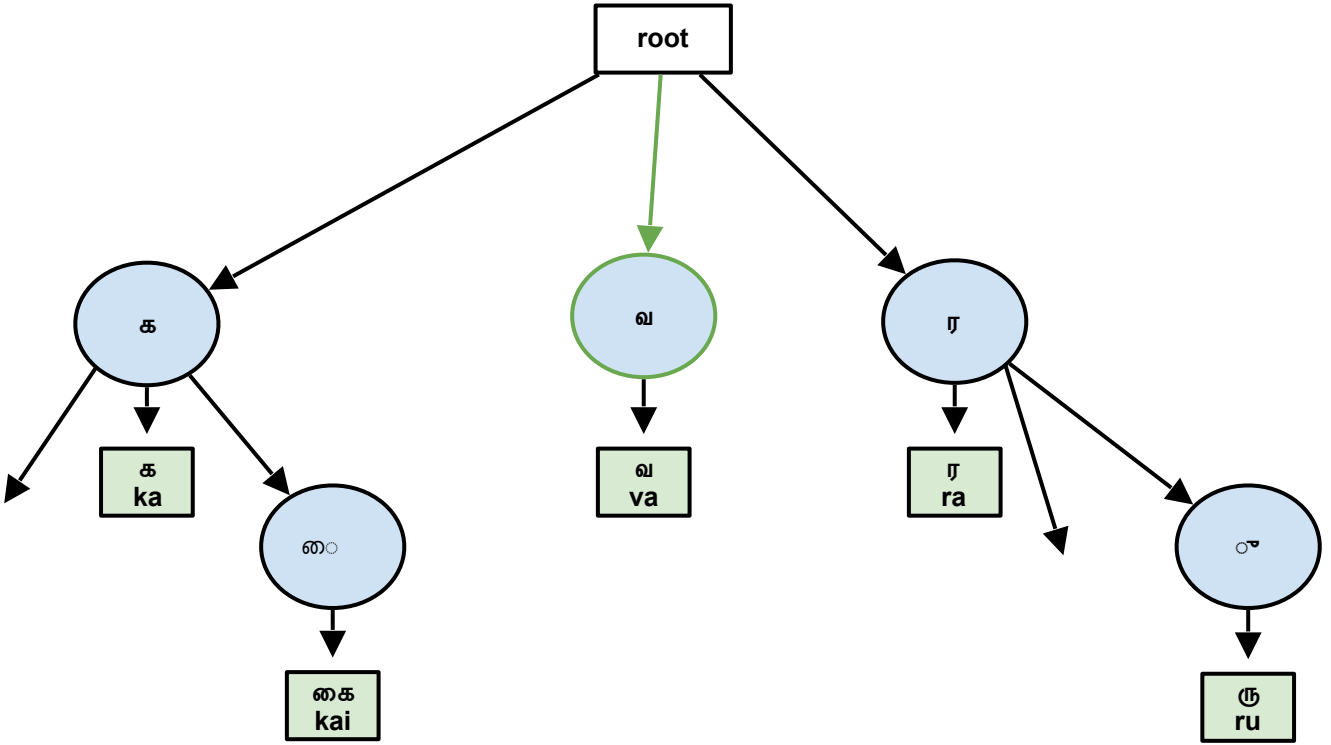
Output:



Prefix Tree Parsing Tamil Letters

Input:
வ, ர, ◌ை, க, ை◌

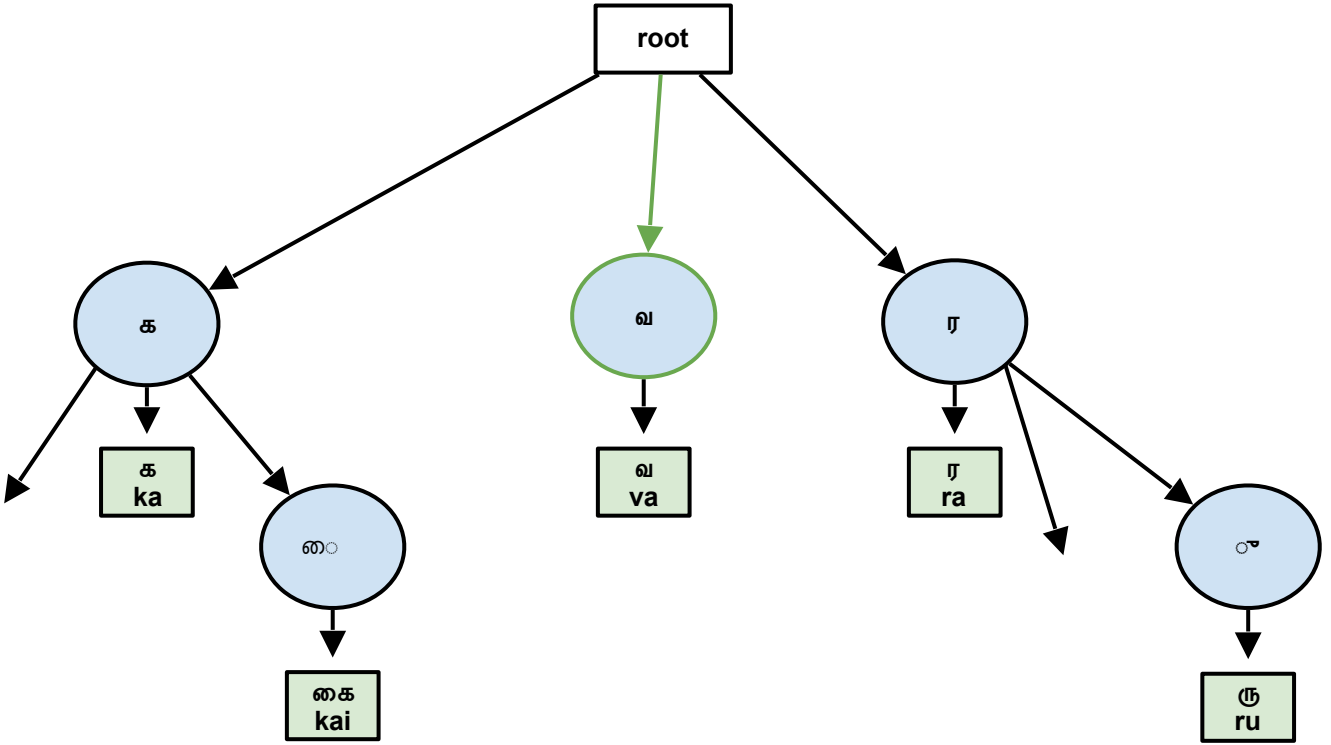
Output:



Prefix Tree Parsing Tamil Letters

Input:
வ, ர, ◌ை, க, ை◌

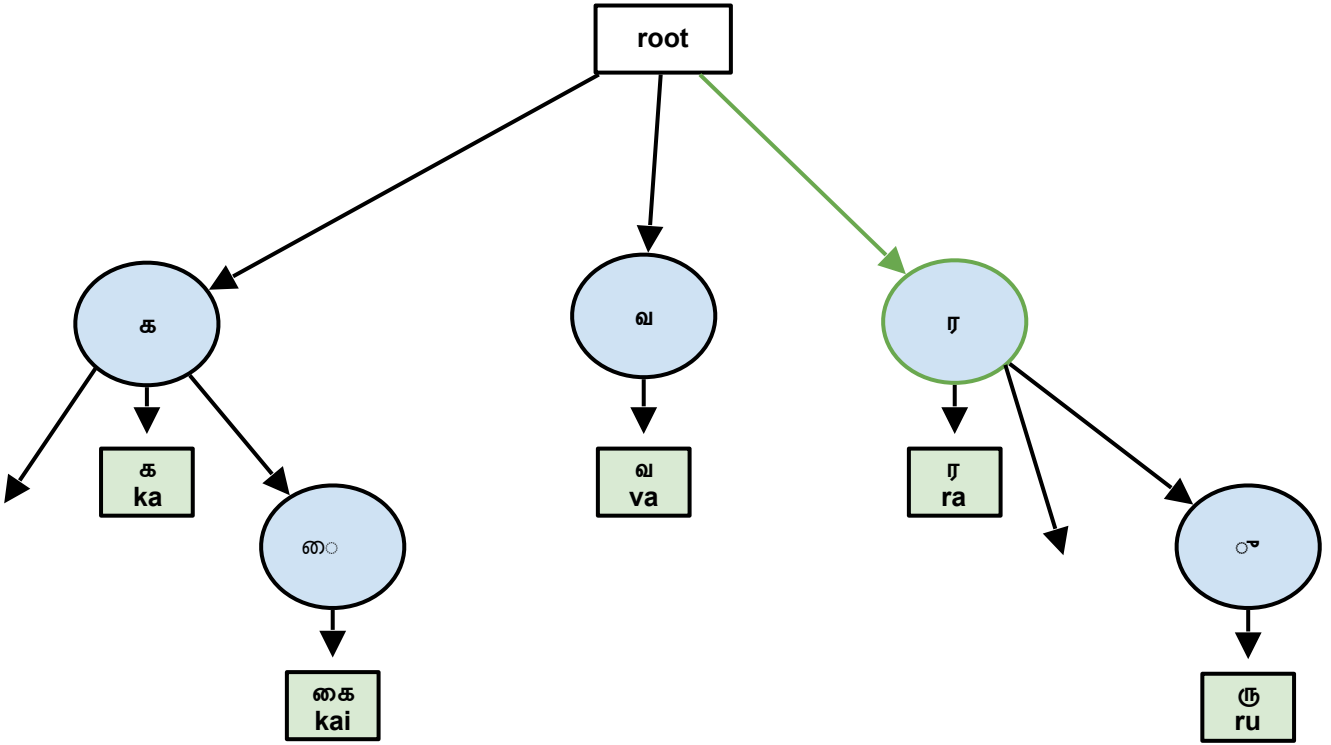
Output:
வ



Prefix Tree Parsing Tamil Letters

Input:
ர, ◌ை, க, ை◌

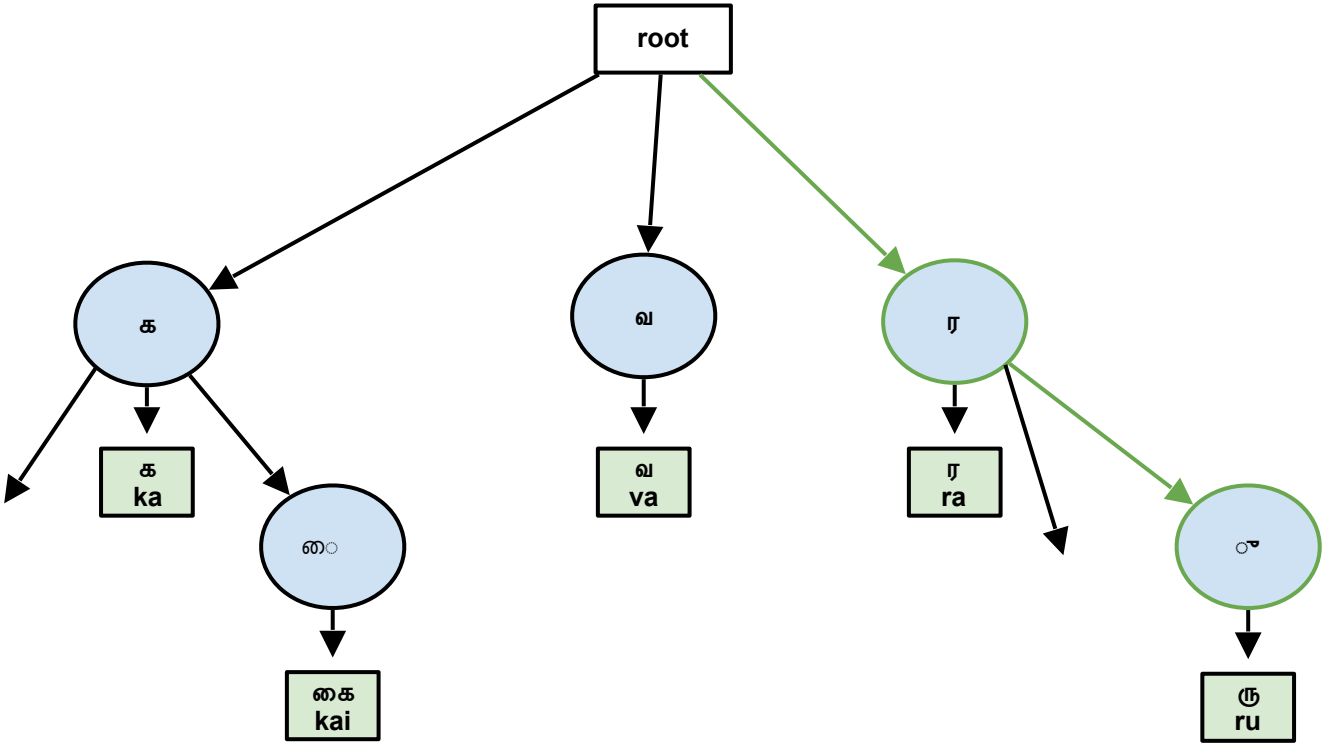
Output:
வ



Prefix Tree Parsing Tamil Letters

Input:
ர, ◌ு, க, ை◌

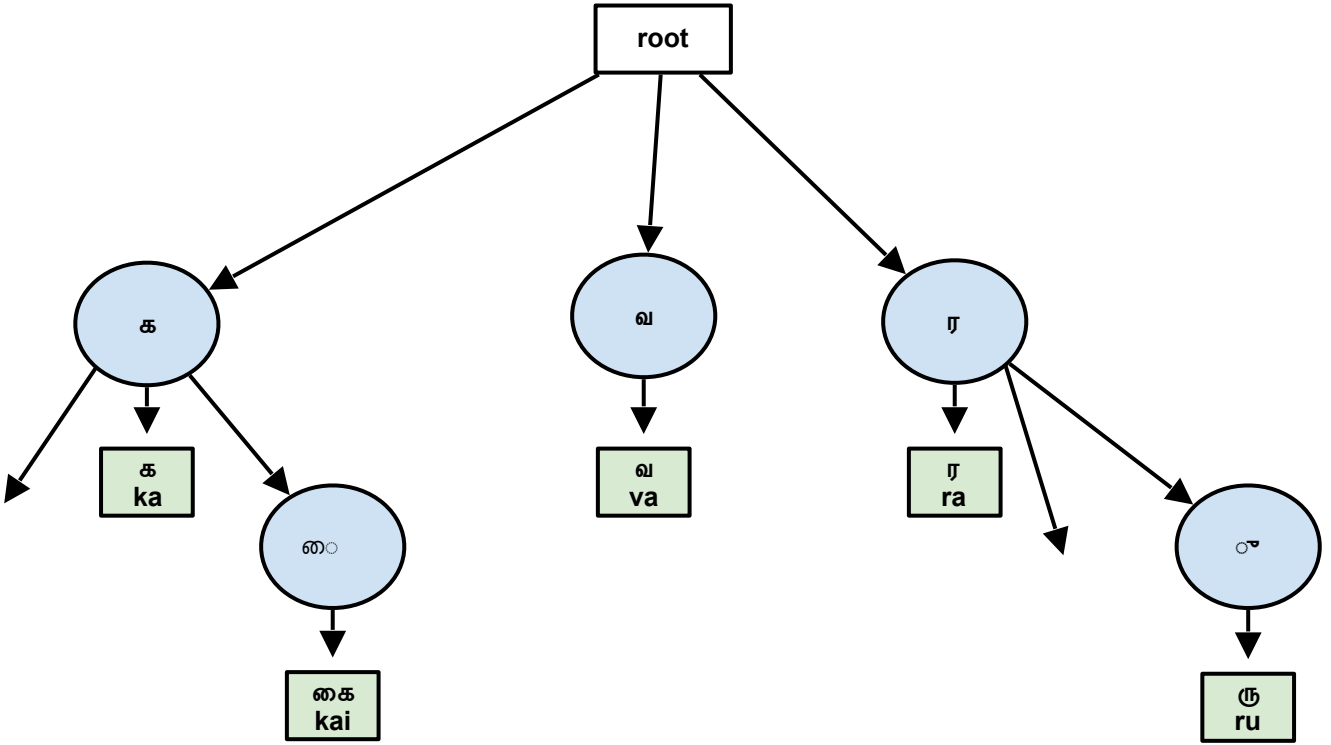
Output:
வ



Prefix Tree Parsing Tamil Letters

Input:
ர, ◌ை, க, ை◌

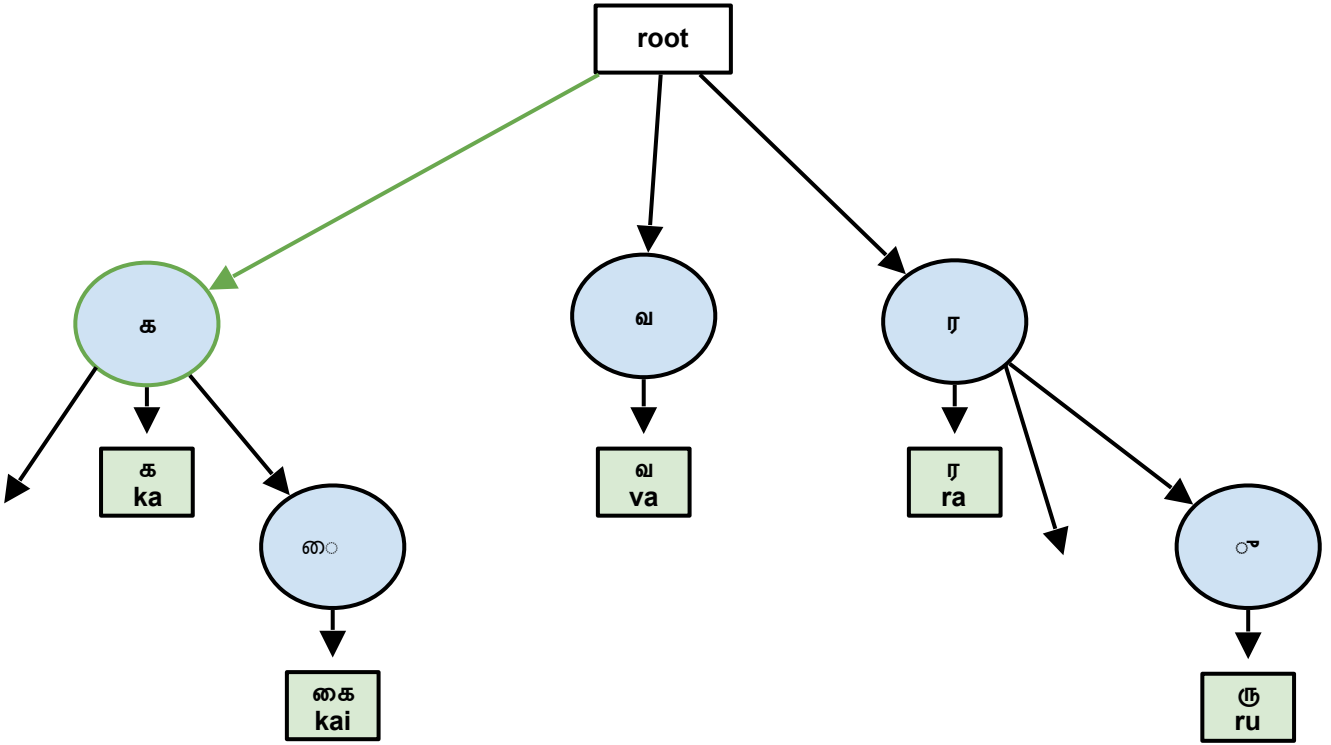
Output:
வ, ரு



Prefix Tree Parsing Tamil Letters

Input:
க, ை

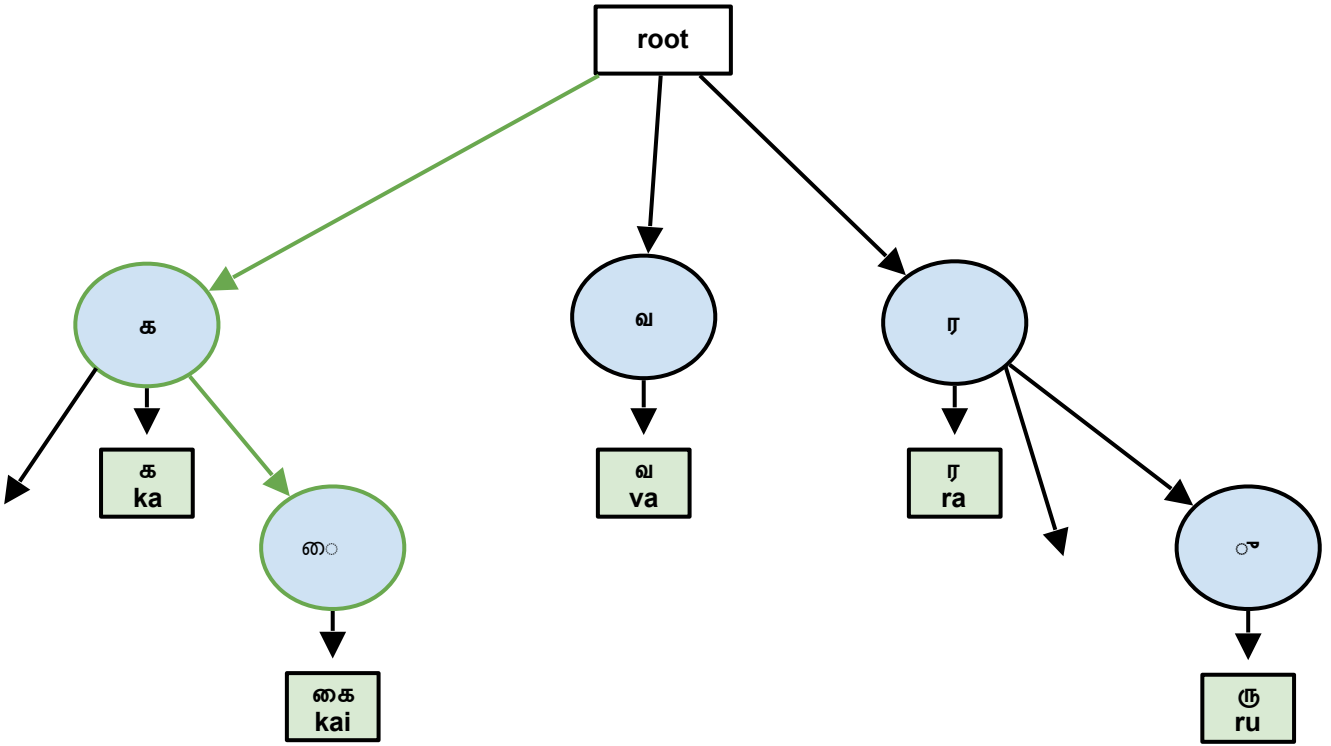
Output:
வ, ரு



Prefix Tree Parsing Tamil Letters

Input:
க, னை

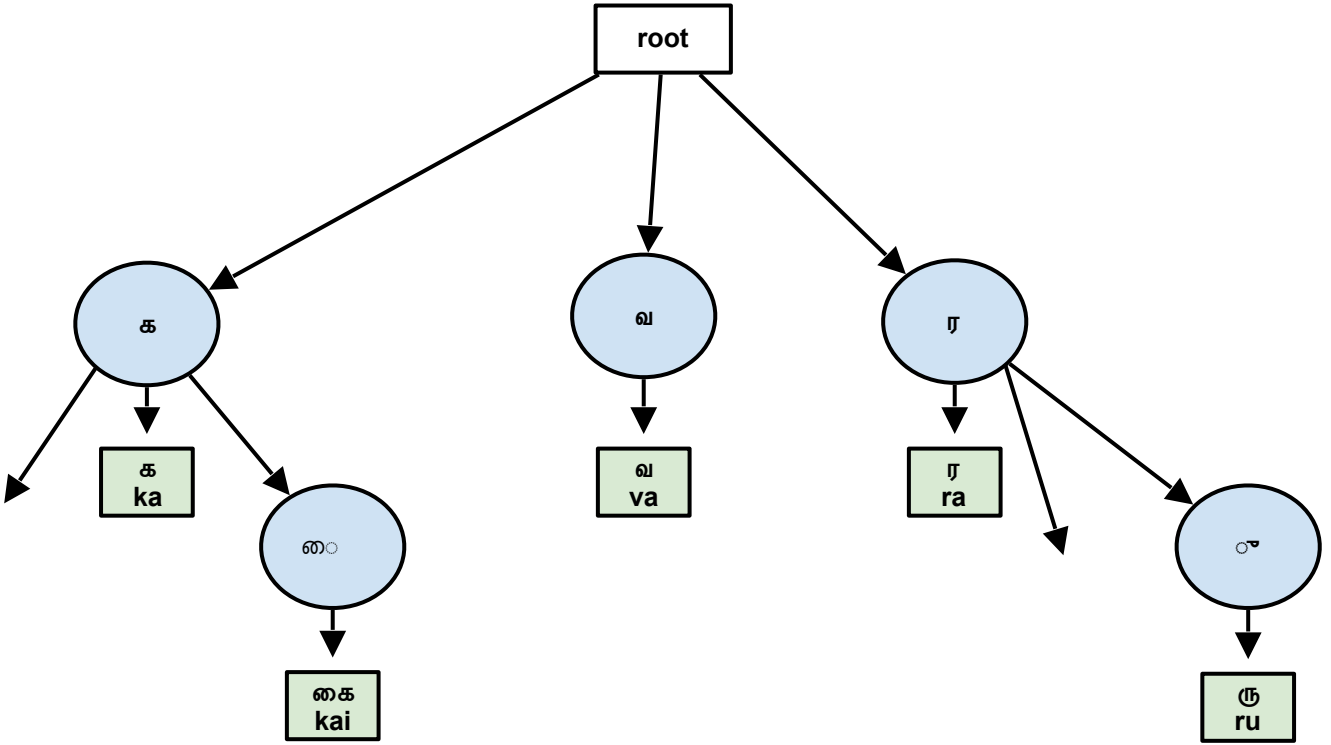
Output:
வ, ரு



Prefix Tree Parsing Tamil Letters

Input:

Output:
வ, ரு, கை



Phonemes in Tamil

- Tamil grammar is based on phonemes (sounds), not how we write
  - Tamil letters sometimes have 1 or 2 sounds
- We should convert text into phonemes for Tamil grammar operations

காடு + இல்
= [க், ஆ, ட், உ] + [இ, ல்]
-> [க், ஆ, ட், ட்] + [இ, ல்]
= [க், ஆ, ட், ட், இ, ல்]
-> [கா, ட், டி, ல்]

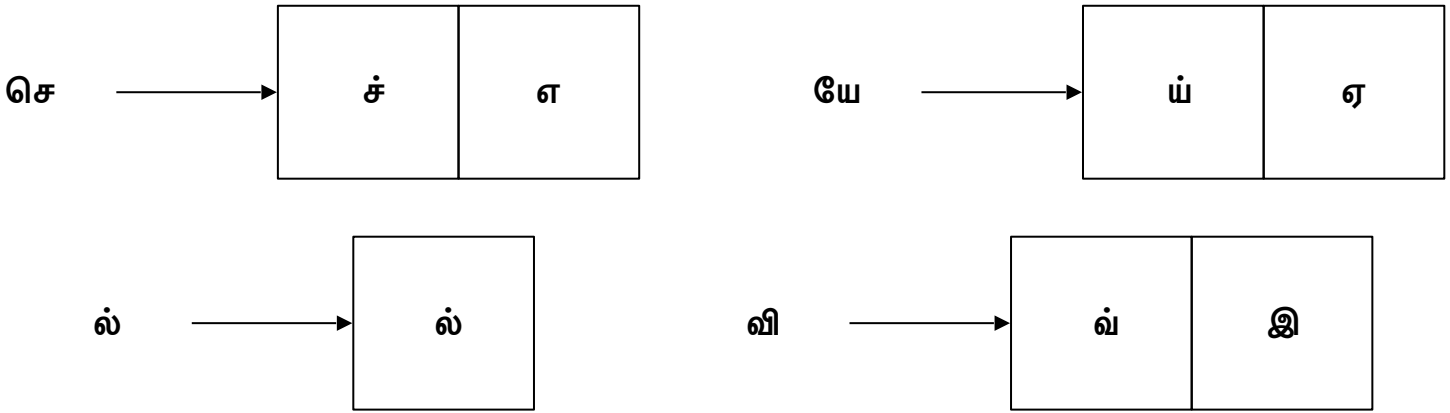
kaadu + il
= [k, aa, d, u] + [i, l]
-> [k, aa, t, t] + [i, l]
= [k, aa, t, t, i, l]
-> [kaa, t, ti, l]

தமிழில் ஒலியன்கள்

- தமிழ் இலக்கணம் ஒலியன் அடிப்படையில் இருக்கும், எழுதும் முறை மூலம் அல்ல
  - ஓர் எழுத்தில் 1 அல்லது 2 ஒலியன்கள் இருக்கும்
- தமிழ் இலக்கணத்துக்கு உரையை ஒலியன்களாகப் பிரிக்கவேண்டும்

Prefix Trees for Tamil Phonemes

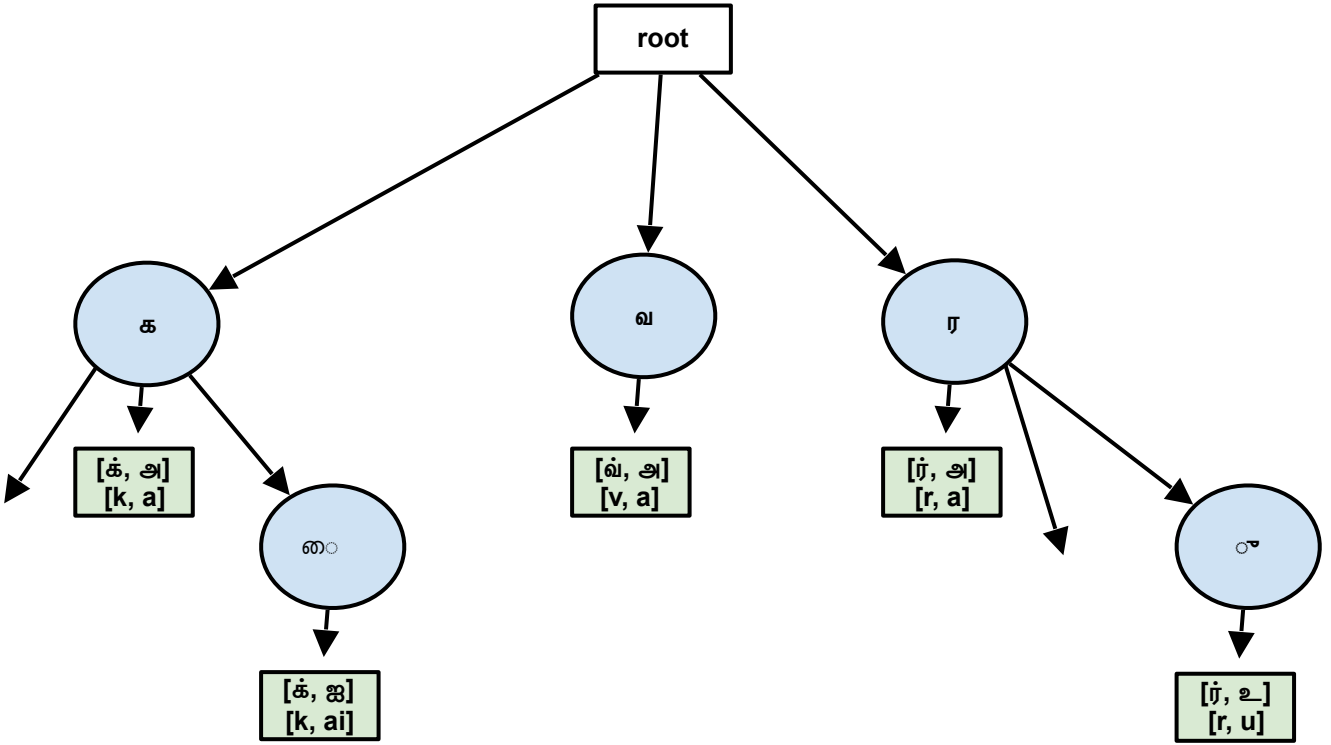
- Make the prefix tree like a map
  - Associate value to each sequence at leaf node in tree
- Associate each Tamil letter's character sequence in tree with the sequence of phonemes
- Also create inverse prefix tree



Prefix Tree Parsing Tamil Phonemes

Input:
வ, ர, ◌ை, க, ை◌

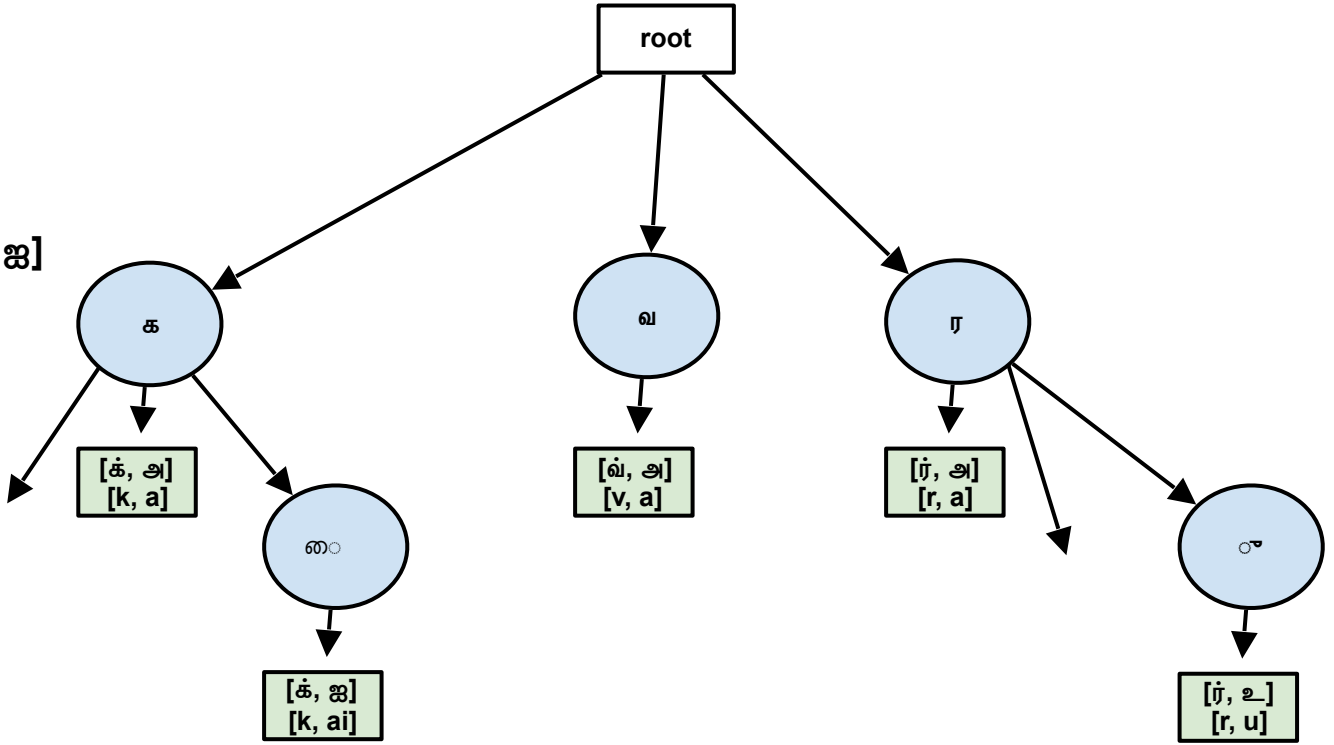
Output:



Prefix Tree Parsing Tamil Phonemes

Input:

Output:
[வ், அ, ர், உ, க், ஐ]



Functions for Tamil Grammar

- Tamil grammar rules can be written as functions of phonemes that returns phonemes

தமிழ் இலக்கணத்-துக்கான செயல்-கூறுகள்

- ஒலியன்க்ளை உள்வாங்கி ஒலியங்களைக் கொடுக்கும் செயல் கூறுகளாகத் தமிழ் இலக்கண விதிகளை எழுதலாம்

```
(வரையறு-செயல்கூறு பன்மை
 "ஒரு சொல்லை அதன் பன்மை வடிவத்தில் ஆக்குதல்
 takes a word and pluralizes it"
 [சொல்]
 (வைத்துக்கொள் [எழுத்துகள் (தொடை->எழுத்துகள் சொல்)]
   (பொறுத்து
    (பின்னொட்டா? சொல் "கள்")                    சொல்
    (= "ம்" (கடைசி எழுத்துகள்))                  (செயல்படுத்து தொடை (தொடு
(கடைசியின்றி எழுத்துகள்) ["ங்கள்"]))
    (மற்றும் (= 1 (எண்ணு எழுத்துகள்))
             (நெடிலா? சொல்))                     (தொடை சொல் "க்கள்")
    (மற்றும் (= 2 (எண்ணு எழுத்துகள்))
             (ஒவ்வொன்றுமா? அடையாளம்
                 (விவரி குறிலா? எழுத்துகள்)))   (தொடை சொல் "க்கள்")
    (மற்றும் (= 2 (எண்ணு எழுத்துகள்))
             (குறிலா? (முதல் எழுத்துகள்))
             (= "ல்" (இரண்டாம் எழுத்துகள்)))    (தொடை (முதல் எழுத்துகள்) "ற்கள்")
    (மற்றும் (= 2 (எண்ணு எழுத்துகள்))
             (குறிலா? (முதல் எழுத்துகள்))
             (= "ள்" (இரண்டாம் எழுத்துகள்)))    (தொடை (முதல் எழுத்துகள்) "ட்கள்")
    :அன்றி                                       (தொடை சொல் "கள்"))))
```

```
(வரையறு-செயல்கூறு சந்தி
 [சொல்1 சொல்2]
 (வைத்துக்கொள் [எழுத்துகள்1 (தொடை->எழுத்துகள் சொல்1)
                 எழுத்துகள்2 (தொடை->எழுத்துகள் சொல்2)
                 ஒலியன்கள்1 (தொடை->ஒலியன்கள் சொல்1)
                 ஒலியன்கள்2 (தொடை->ஒலியன்கள் சொல்2)
                 சொ1-கஒ (கடைசி ஒலியன்கள்1)
                 சொ2-முஒ (முதல் ஒலியன்கள்2)]
   (பொறுத்து
    (மற்றும் (உயிரெழுத்தா? சொ2-முஒ)
             (பெறு #{"இ" "ஈ" "ஏ" "ஐ"} சொ1-கஒ))
    (செயல்படுத்து தொடை சொல்1 (ஒலியன்கள->எழுத்து ["ய்" சொ2-முஒ])
(மீதி சொல்2))
    (மற்றும் (உயிரெழுத்தா? சொ2-முஒ)
             (பெறு #{"அ" "ஆ" "ஊ" "ஒ" "ஓ" "ஔ"} சொ1-கஒ))
    (செயல்படுத்து தொடை சொல்1 (ஒலியன்கள->எழுத்து ["வ்" சொ2-முஒ])
(மீதி சொல்2))
    (மற்றும் (உயிரெழுத்தா? சொ2-முஒ)
             (= "உ" சொ1-கஒ)
             (= 2 (எண்ணு எழுத்துகள்1))
             (ஒவ்வொன்றுமா? குறிலா? எழுத்துகள்1))
    (செயல்படுத்து தொடை சொல்1 (ஒலியன்கள->எழுத்து ["வ்" சொ2-முஒ])
(மீதி சொல்2))
    (மற்றும் (உயிரெழுத்தா? சொ2-முஒ)
             (= "உ" சொ1-கஒ)
             (அன்று (மற்றும் (= 2 (எண்ணு எழுத்துகள்1))
                              (ஒவ்வொன்றுமா? குறிலா? எழுத்துகள்1))))
    (செயல்படுத்து தொடை (தொடு (கடைசியின்றி எழுத்துகள்1)
(ஒலியன்கள->எழுத்து [(கடைசி (கடைசியின்றி ஒலியன்கள்1)) சொ2-முஒ])
(மீதி சொல்2)))
    (மற்றும் (உயிரெழுத்தா? சொ2-முஒ)
             (= 2 (எண்ணு எழுத்துகள்1))
             (குறிலா? (முதல் எழுத்துகள்1))
             (மெய்யெழுத்தா? (இரண்டாம் எழுத்துகள்1)))
    (செயல்படுத்து தொடை (தொடு சொல்1 [(ஒலியன்கள->எழுத்து [சொ1-
கஒ சொ2-முஒ])] (மீதி சொல்2)))
    (மற்றும் (உயிரெழுத்தா? சொ2-முஒ)
             (மெய்யெழுத்தா? சொ1-கஒ))
    (செயல்படுத்து தொடை (தொடு (கடைசியின்றி எழுத்துகள்1)
[(ஒலியன்கள->எழுத்து [சொ1-கஒ சொ2-முஒ])] (மீதி சொல்2)))
    :அன்றி
    (தொடை சொல்1 சொல்2))))
```

Demo
செய்துகாட்டுதல்

- clj-thamil.மொழியியல்/
  - தொடை->எழுத்துகள்
  - தொடை->ஒலியன்கள்
  - சந்தி
  - பன்மை
  - வேற்றுமை
- clj-thamil.format/
  - phonemes->str
- clj-thamil.forrmat.convert/
  - தமிழ்->bamini
  - bamini->தமிழ்
  - தமிழ்->romanized
  - romanized->தமிழ்

நன்றி
Thank you