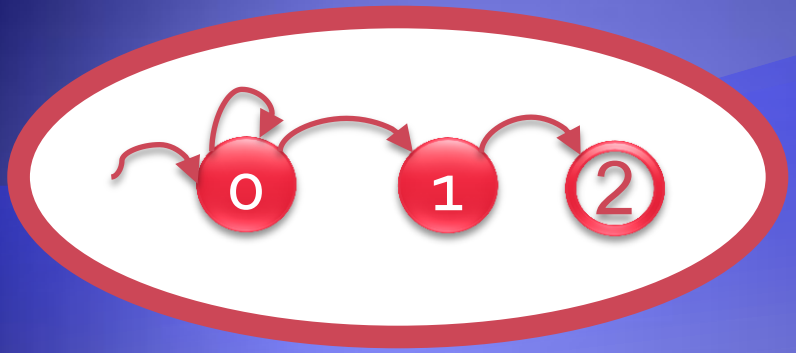
OPEN SOURCE NATURAL LANGUAGE PROCESSING

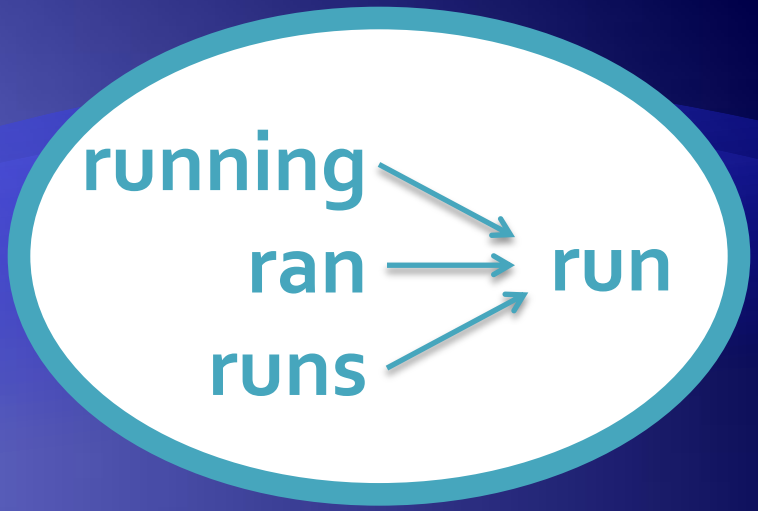
Kara Greenfield

Sarah Judd

Objectives



Finite State Machines
from aff / dic



Stemming

Spell Checking

Open Source

aff / dic

suffix or prefix

Append

SFX	G	e	ing	e
SFX	G	0	ing	[^e]
SFX	D	0	d	e
SFX	D	y	ied	[^aeiou]y
SFX	D	0	ed	[^ey]
SFX	D	0	ed	[aeiou]y

Flag

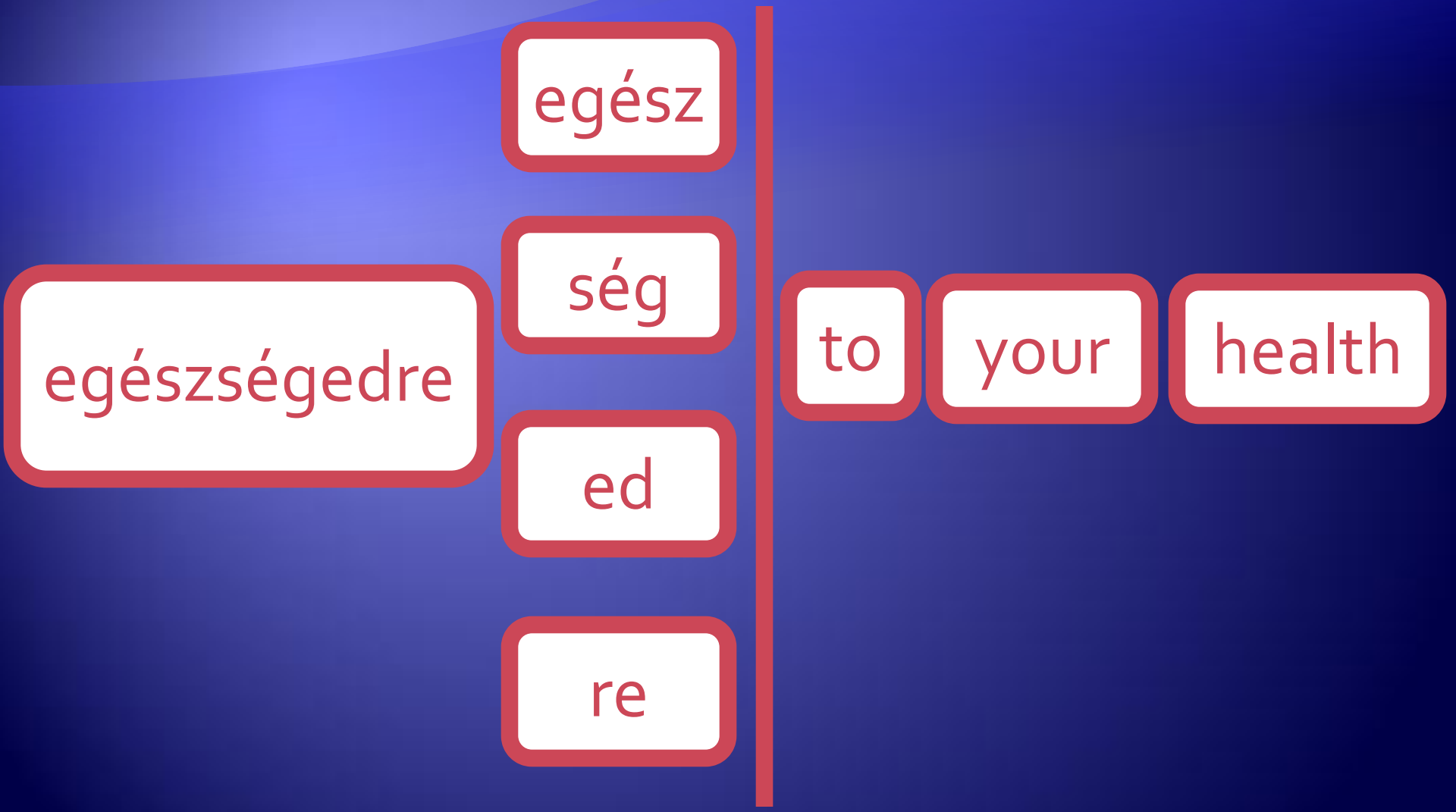
Strip

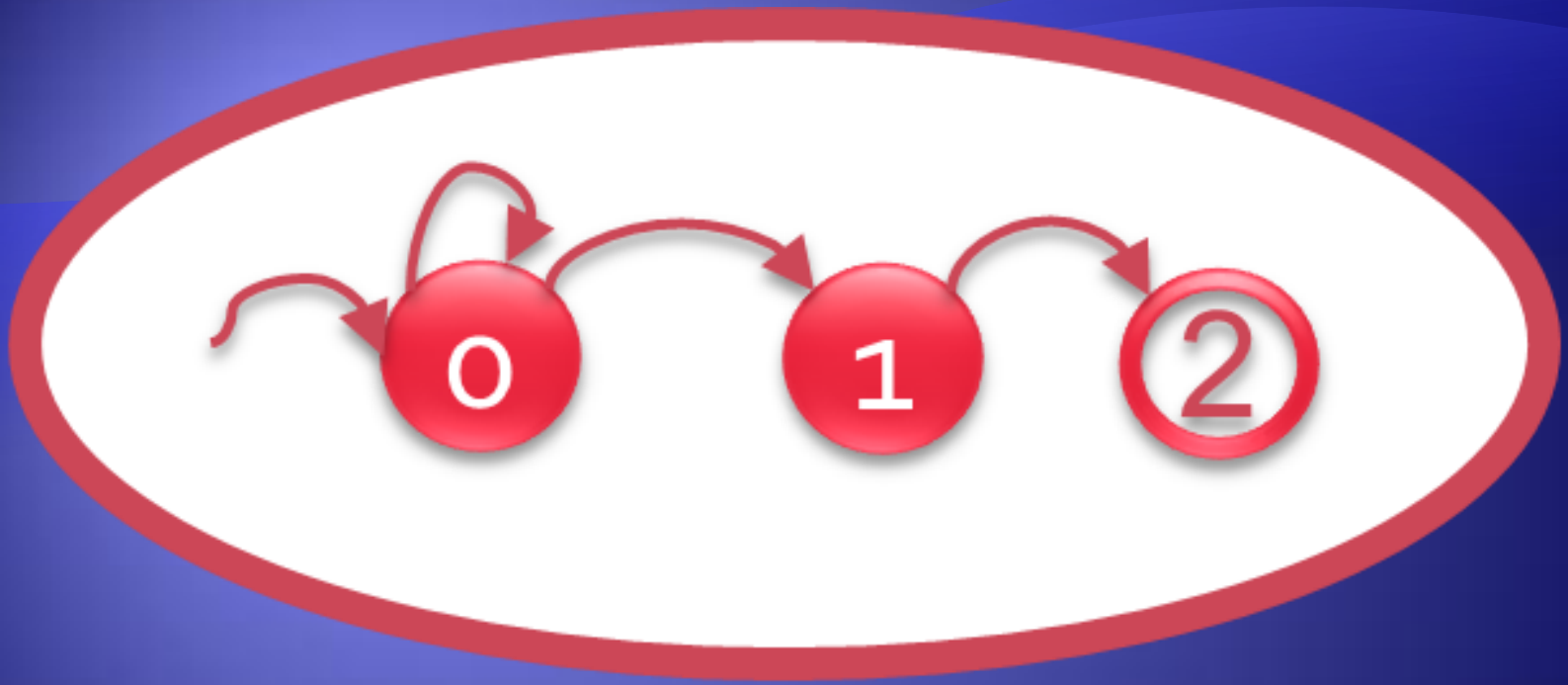
Condition

play/GD
make/G
...

making

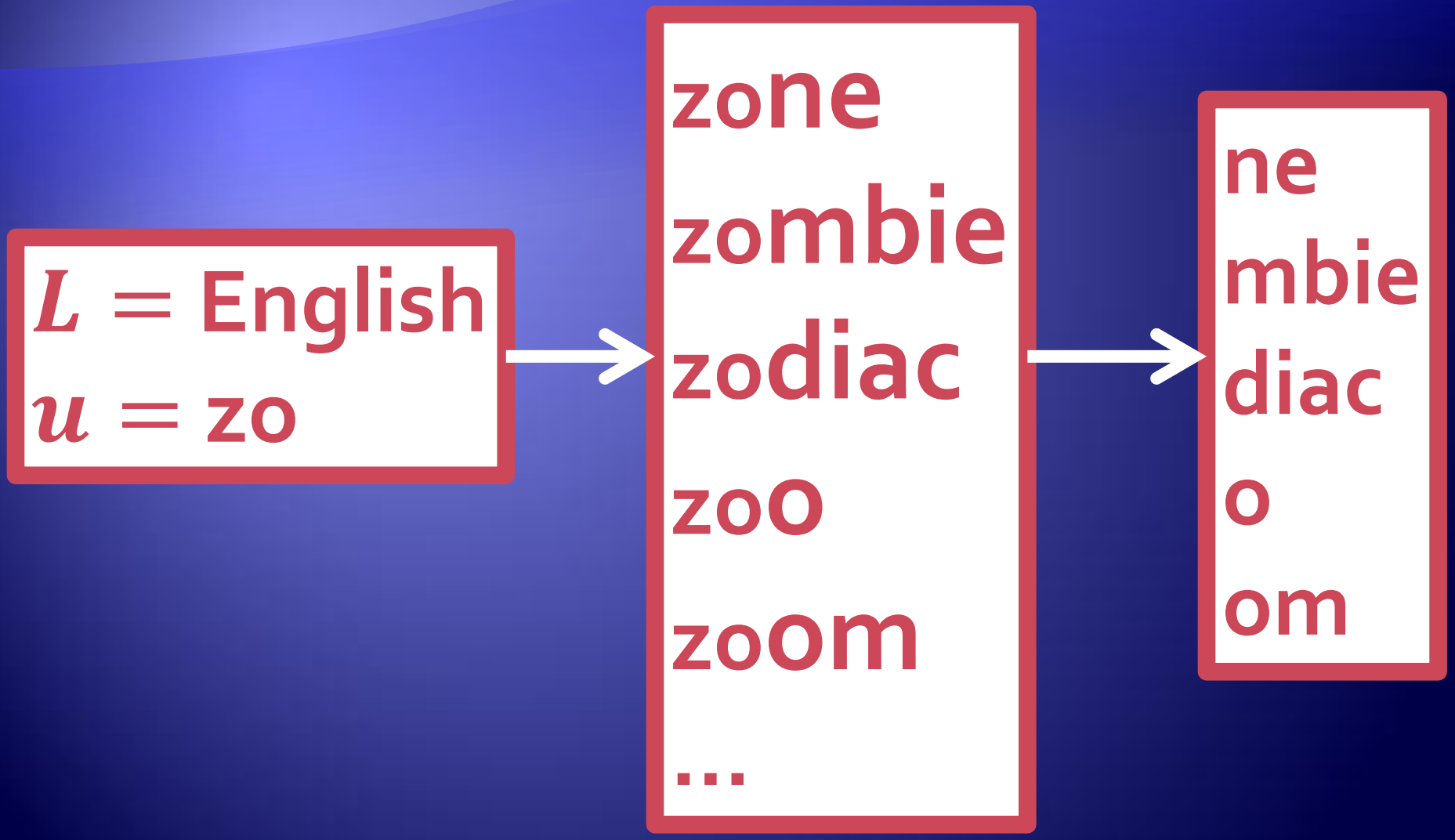
Complex Morphology





Finite State Machines from aff / dic

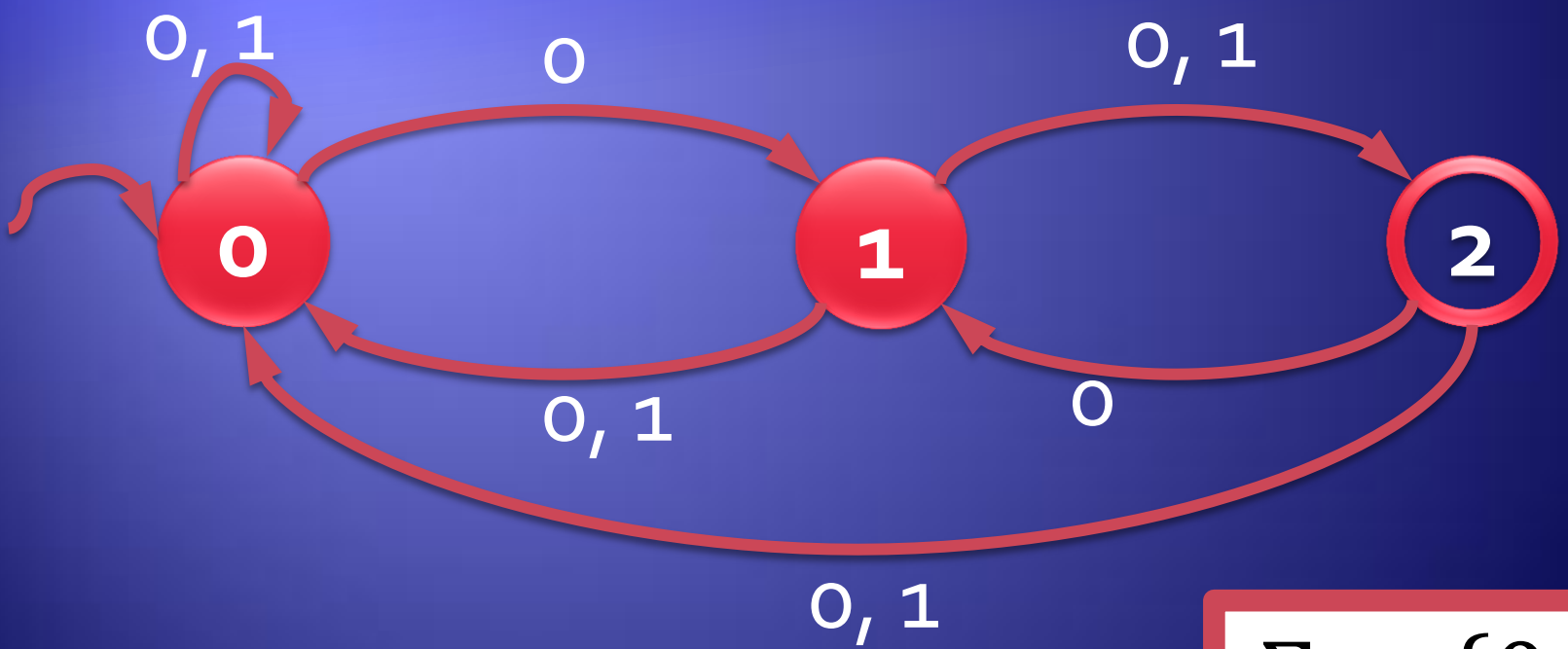
Residual Language



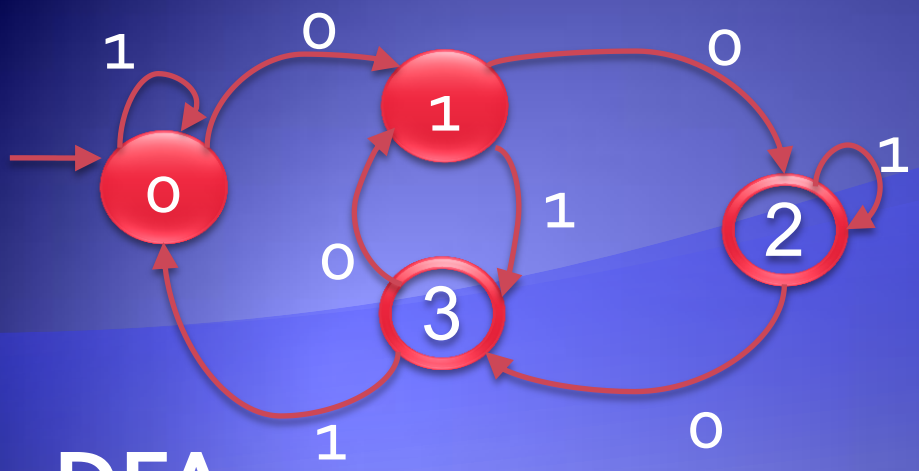
Residual Language

- Let L be a language over Σ^*
- Let $u \in \Sigma^*$
- $u^{-1}L = \{v \in \Sigma^* \mid uv \in L\}$

Residual Finite State Automata



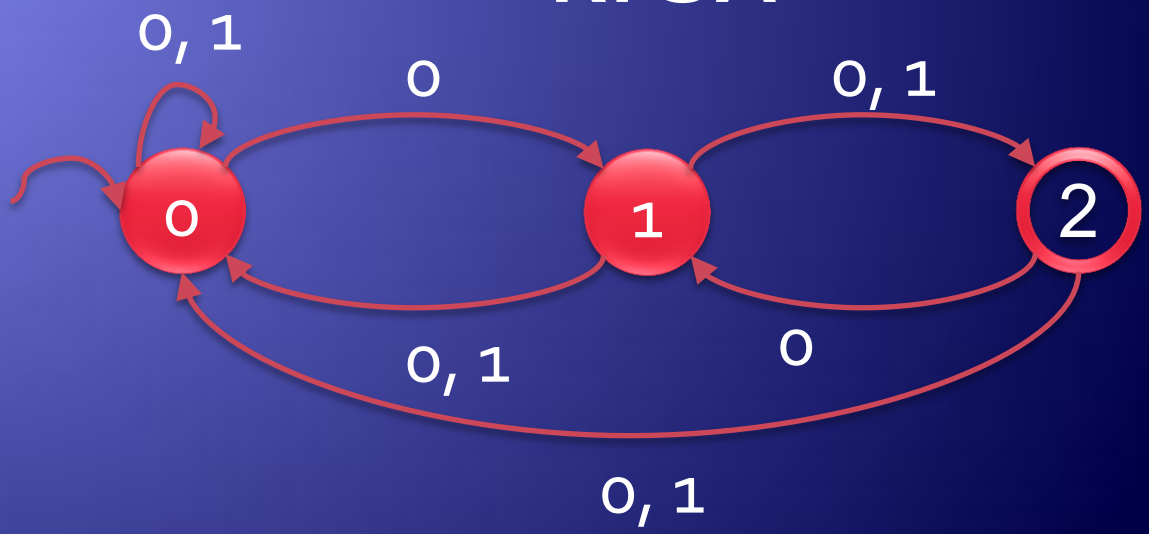
$$\Sigma = \{0, 1\}$$
$$L = \Sigma^* \, 0 \, \Sigma$$



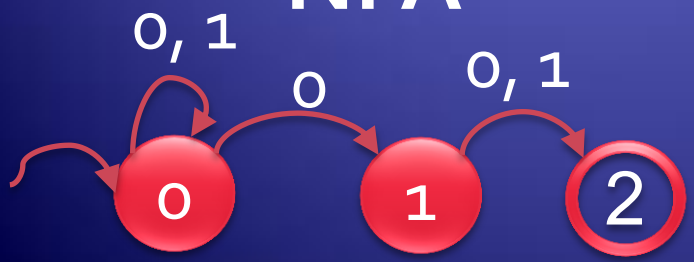
DFA

$$\Sigma = \{0, 1\}$$
$$L = \Sigma^* \, 0 \, \Sigma$$

RFSA



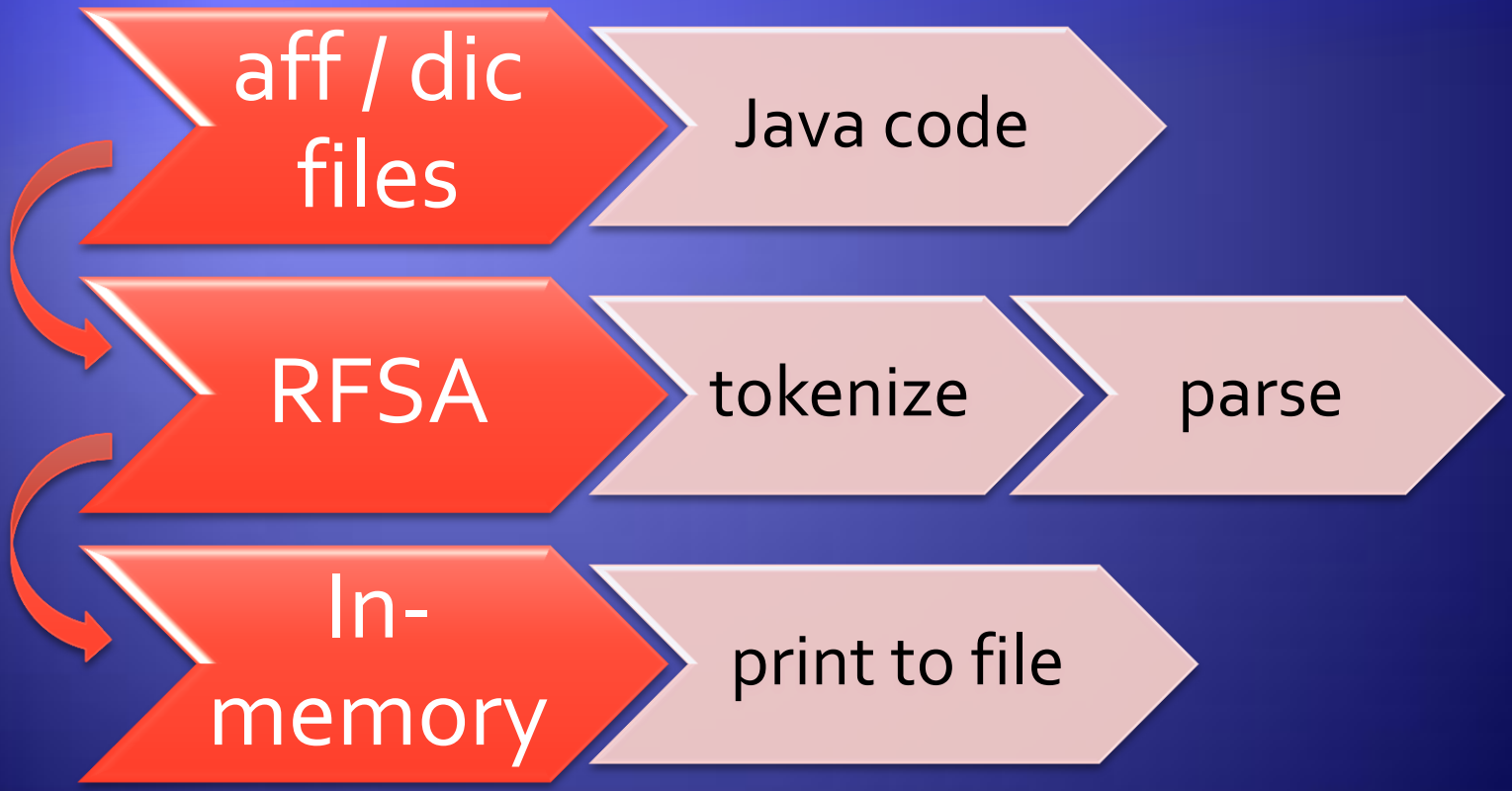
NFA



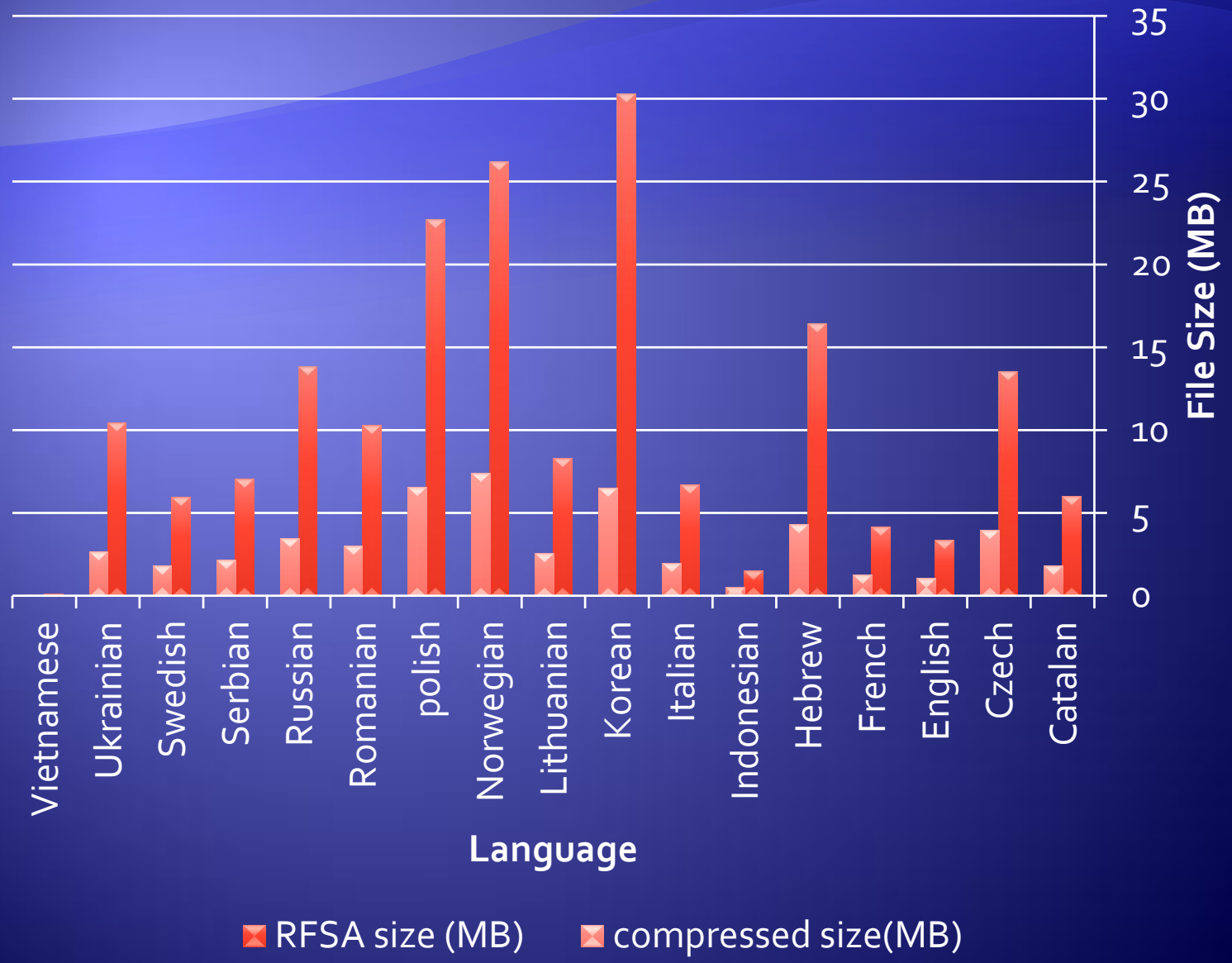
aff/dic → RFSA


aff / dic
NFA words
DFA
Minimize
Compress
Letterize
Minimize
Compress
RFSA

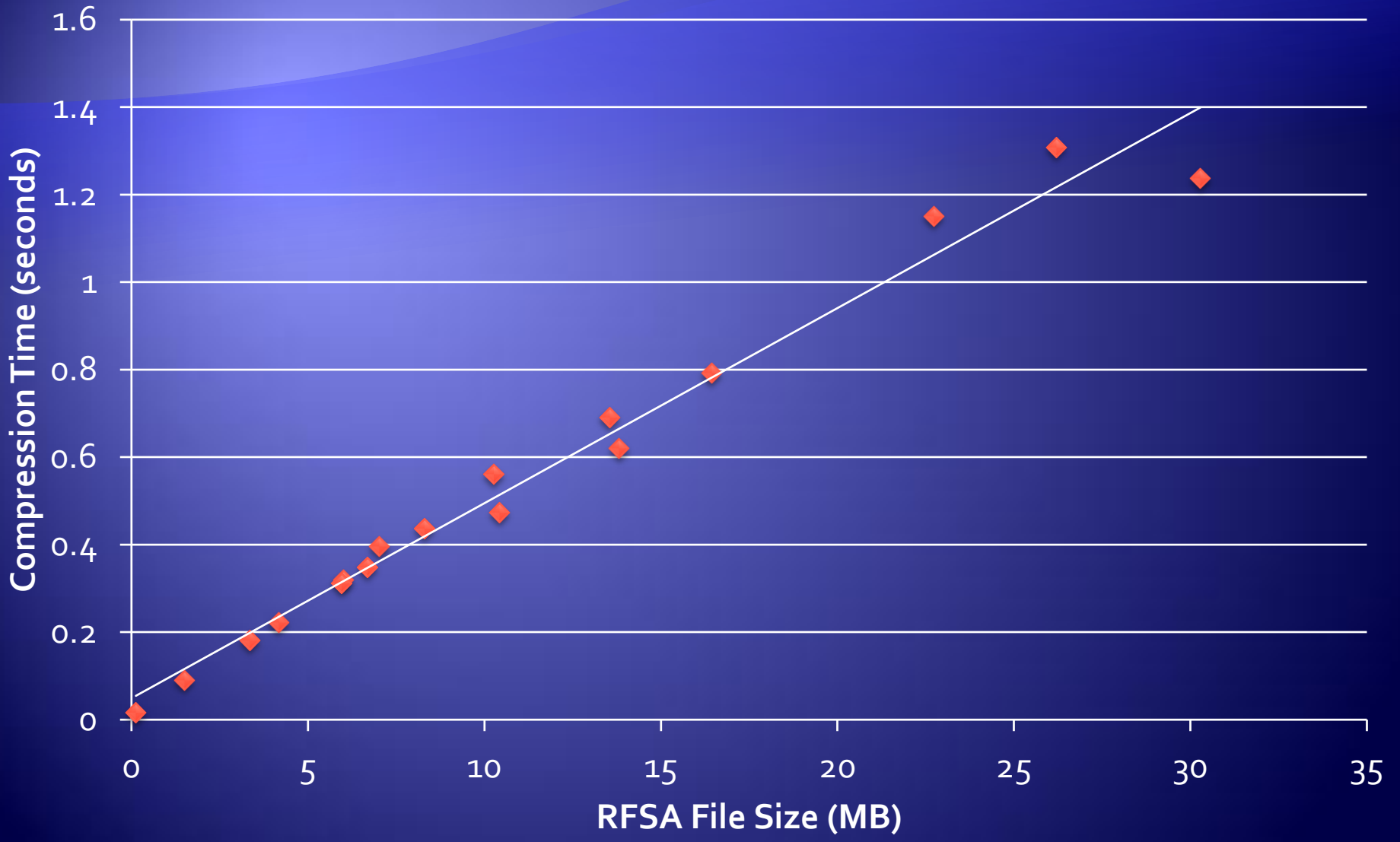
RFSA compression



Language Compression

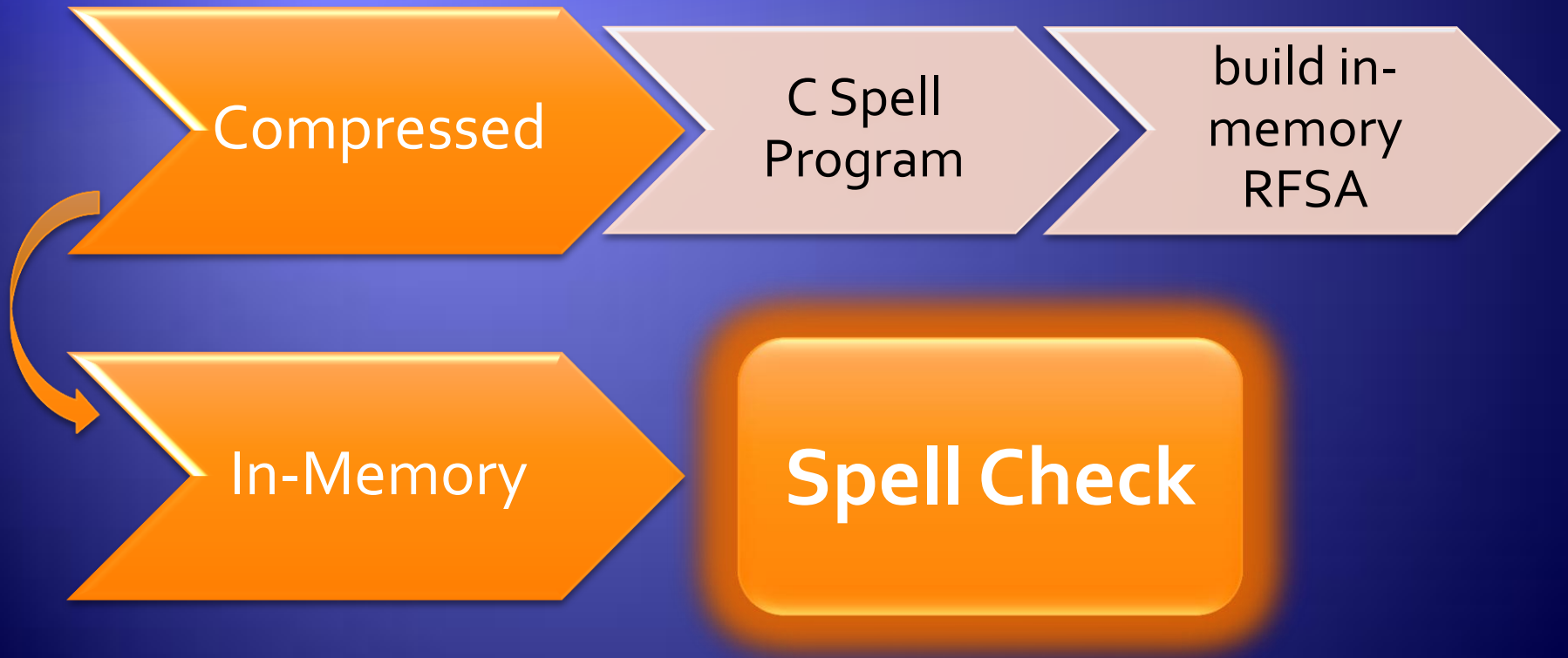


Language Compression

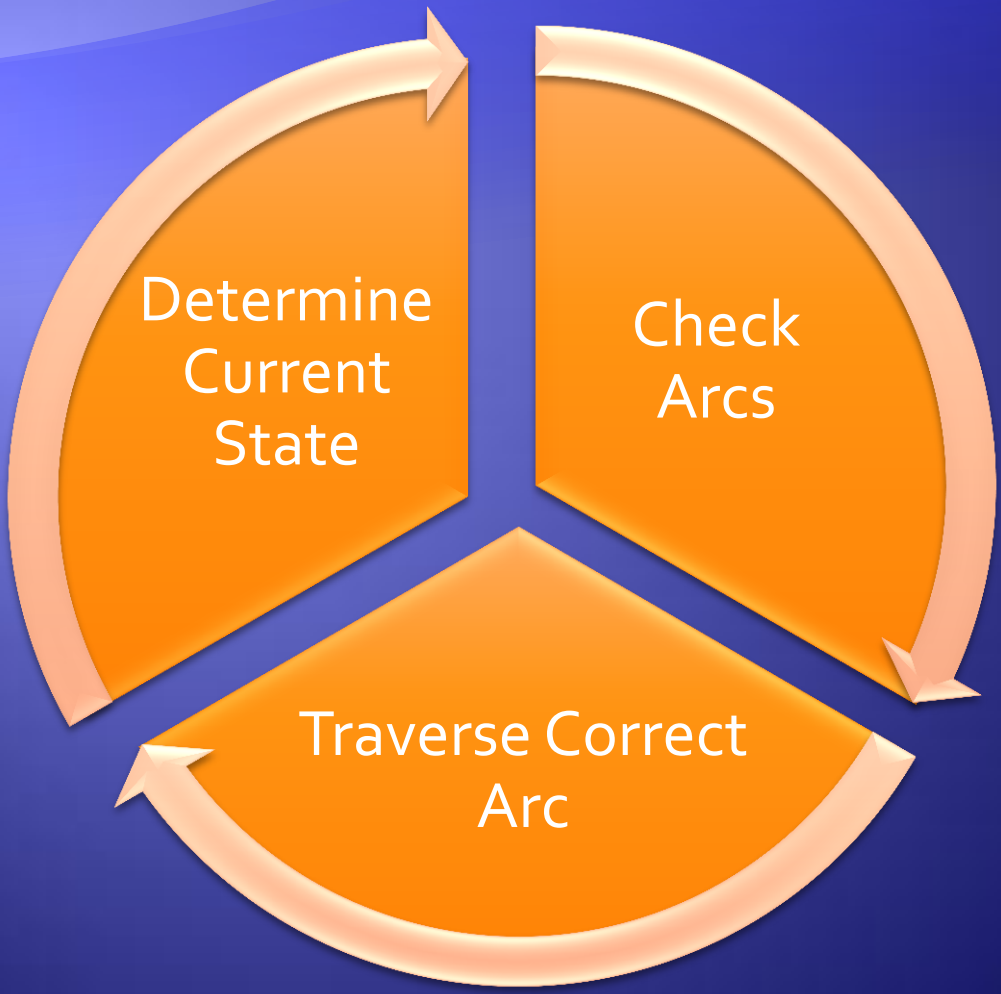


Spell Checking

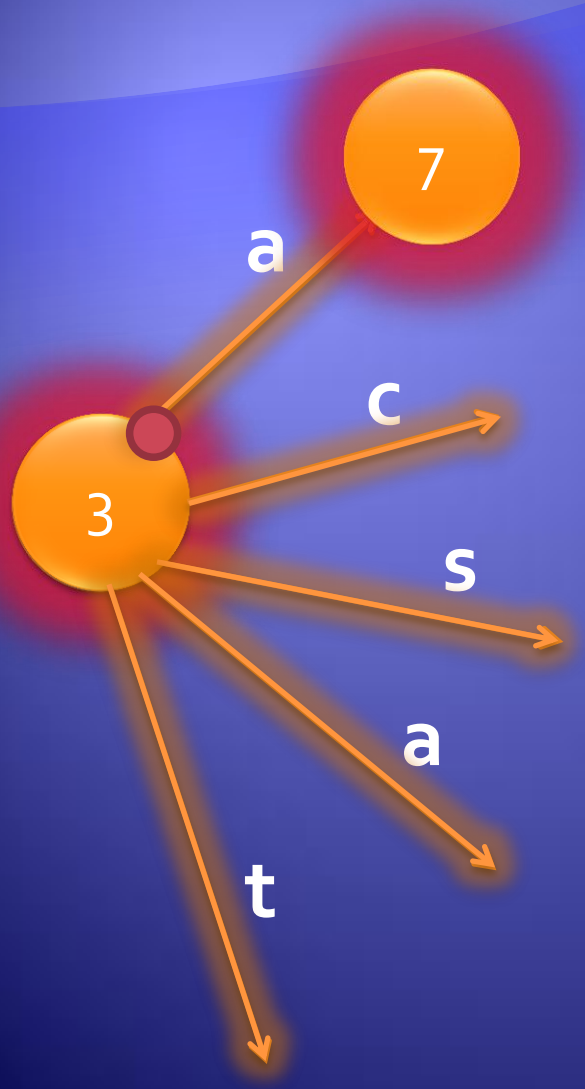
Spell Checking Process



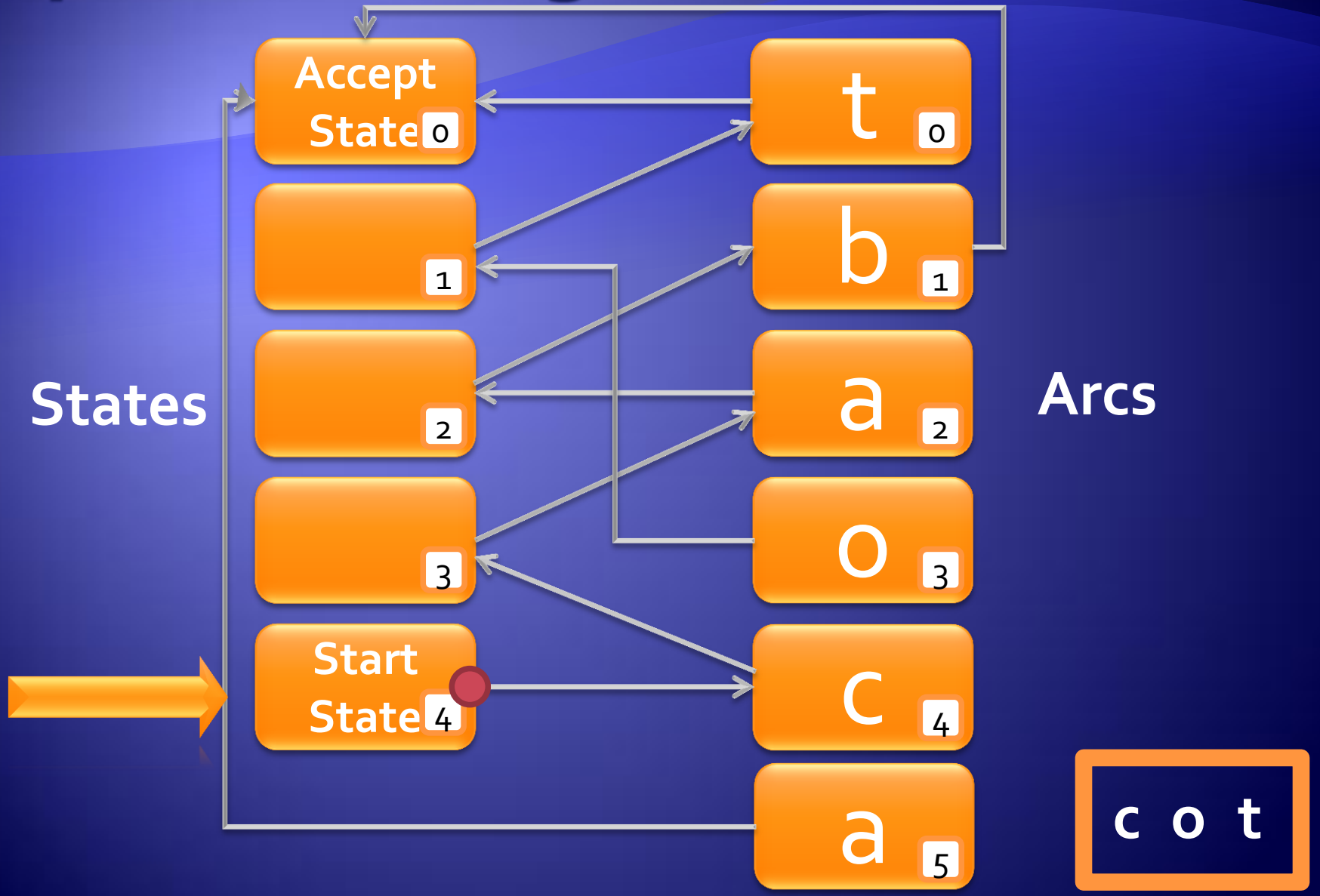
Spell Checking



Spell Checking



Spell Checking

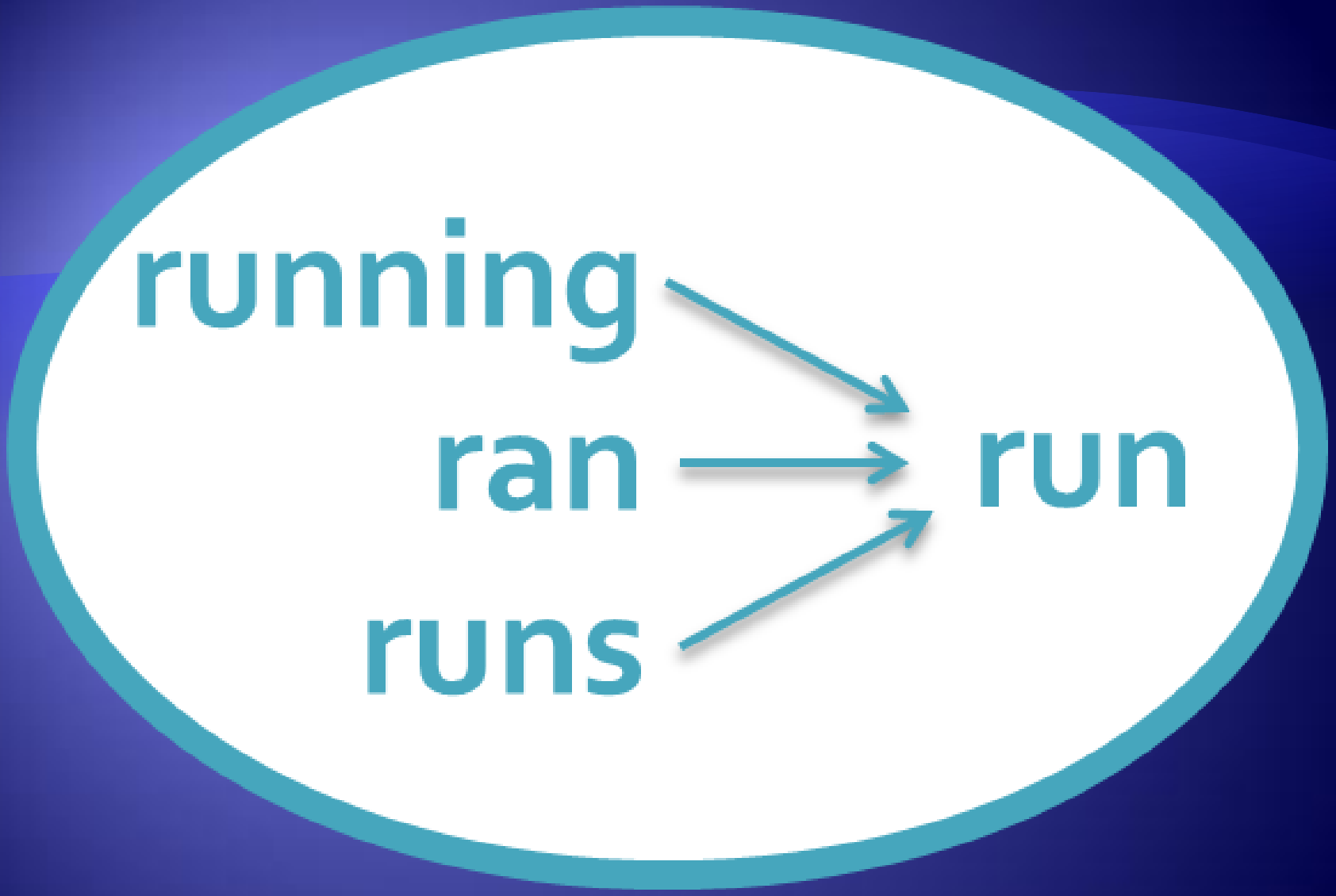


Spell Checking Results

→ 92%

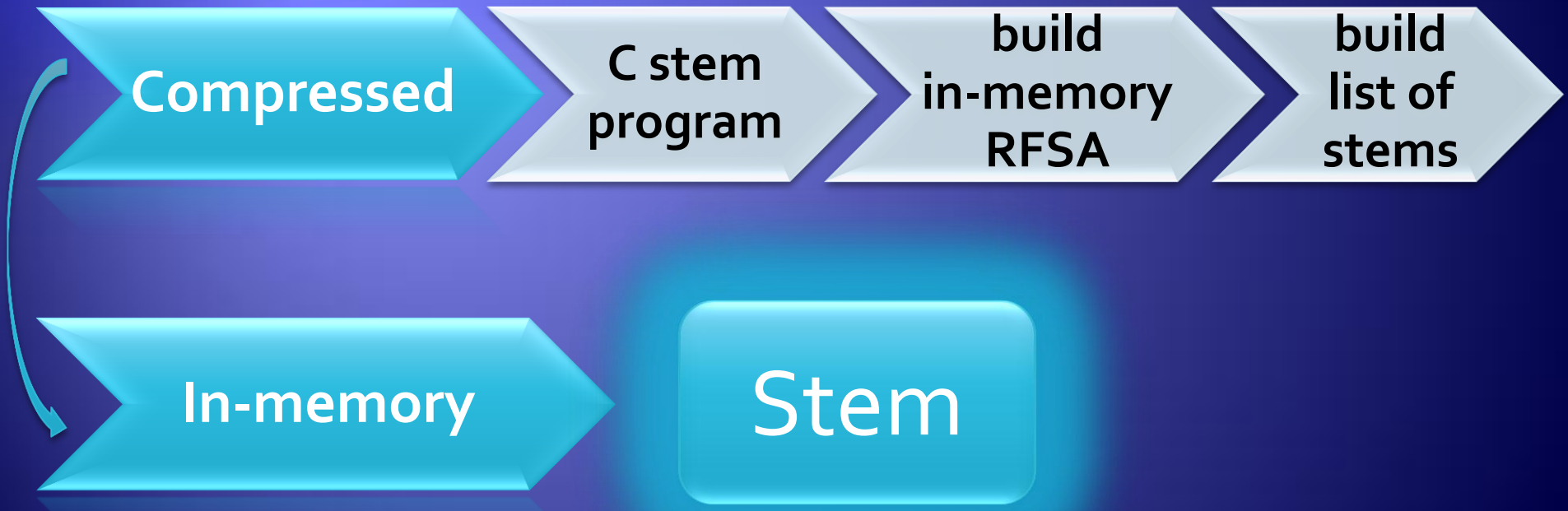
Google → 87%

- non-English
- non-standard abbreviations
- spelling errors
- uncommon proper nouns
- missing from aff/dic

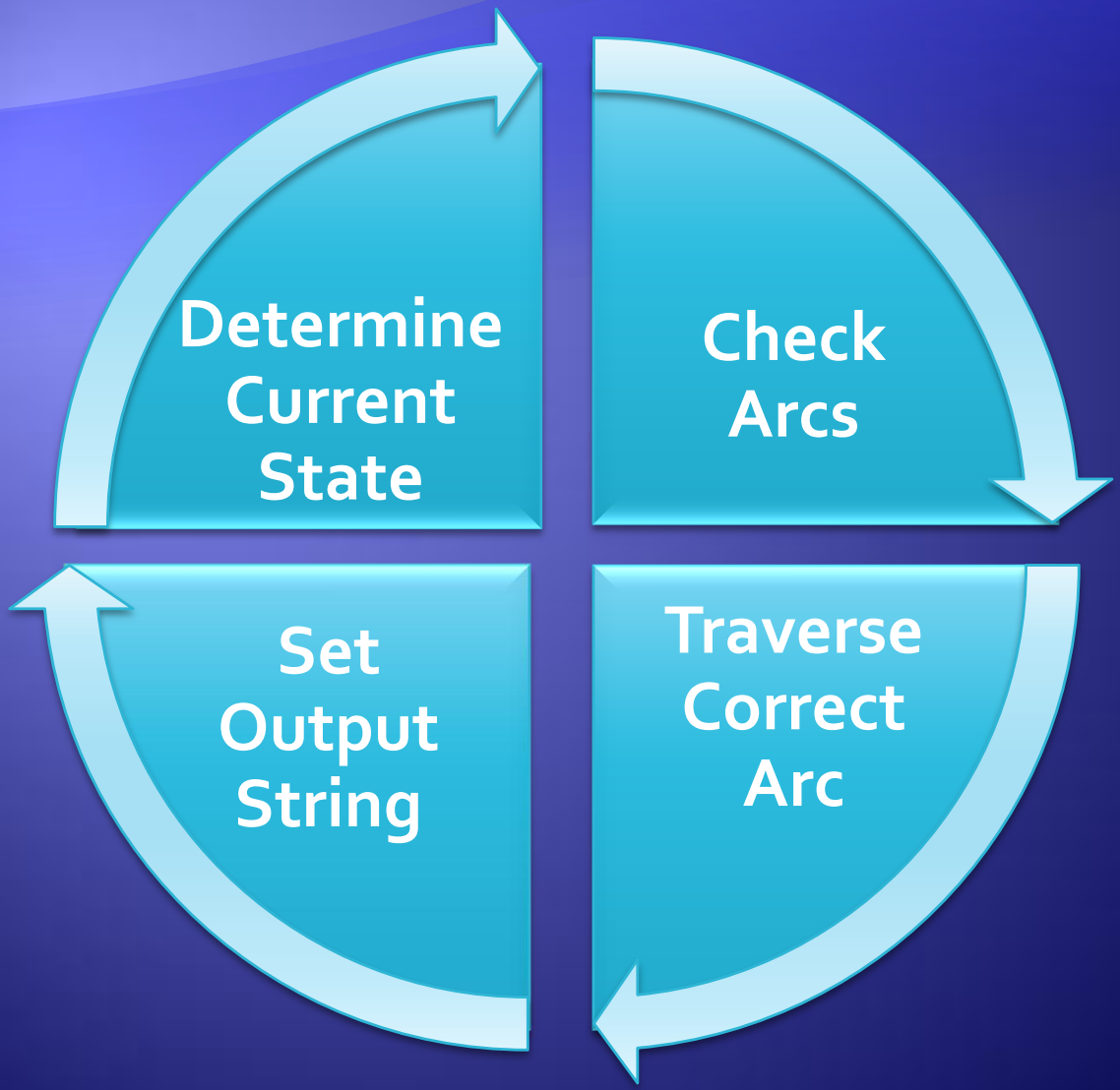


Stemming

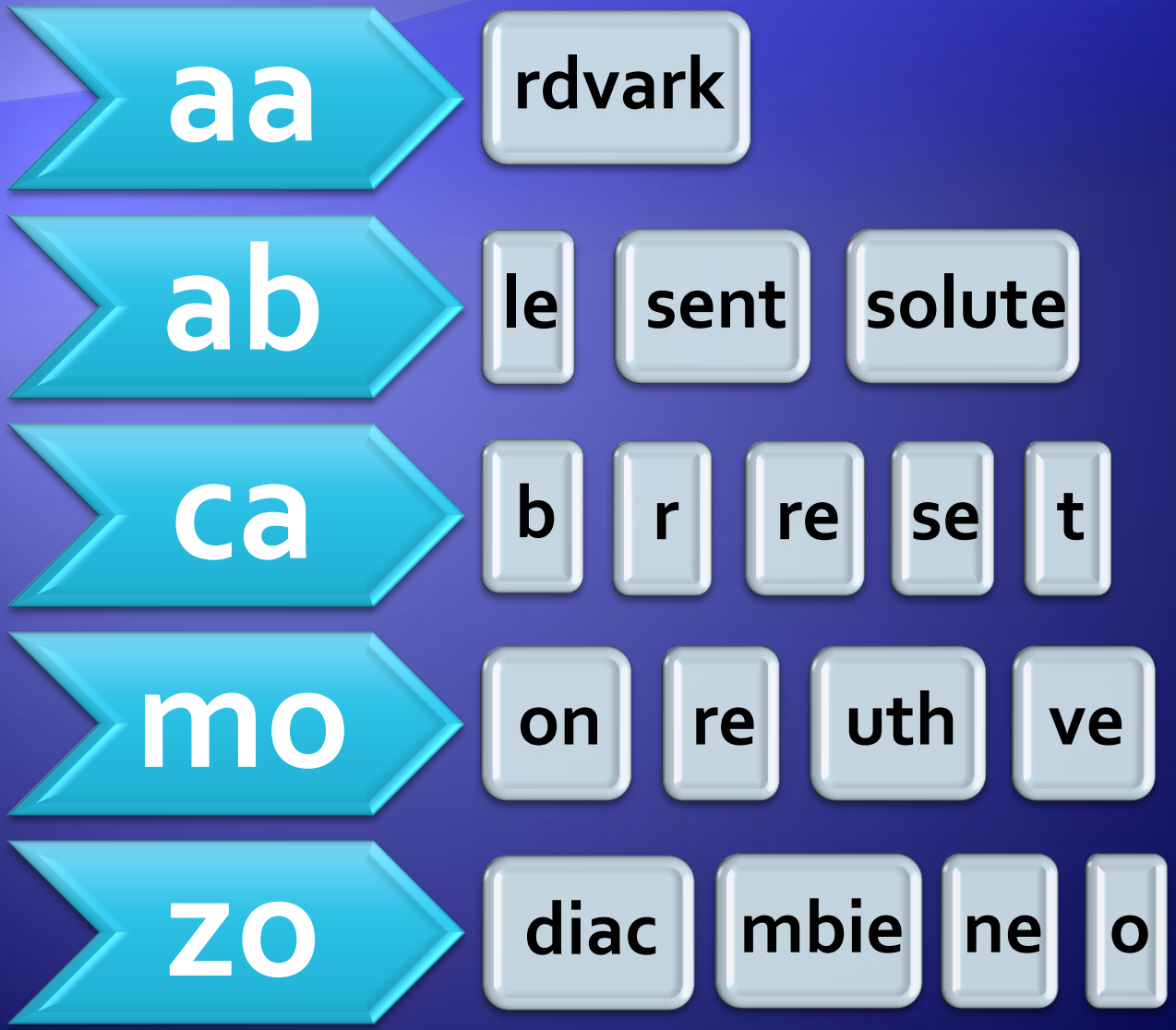
Stemming Process



Stemming

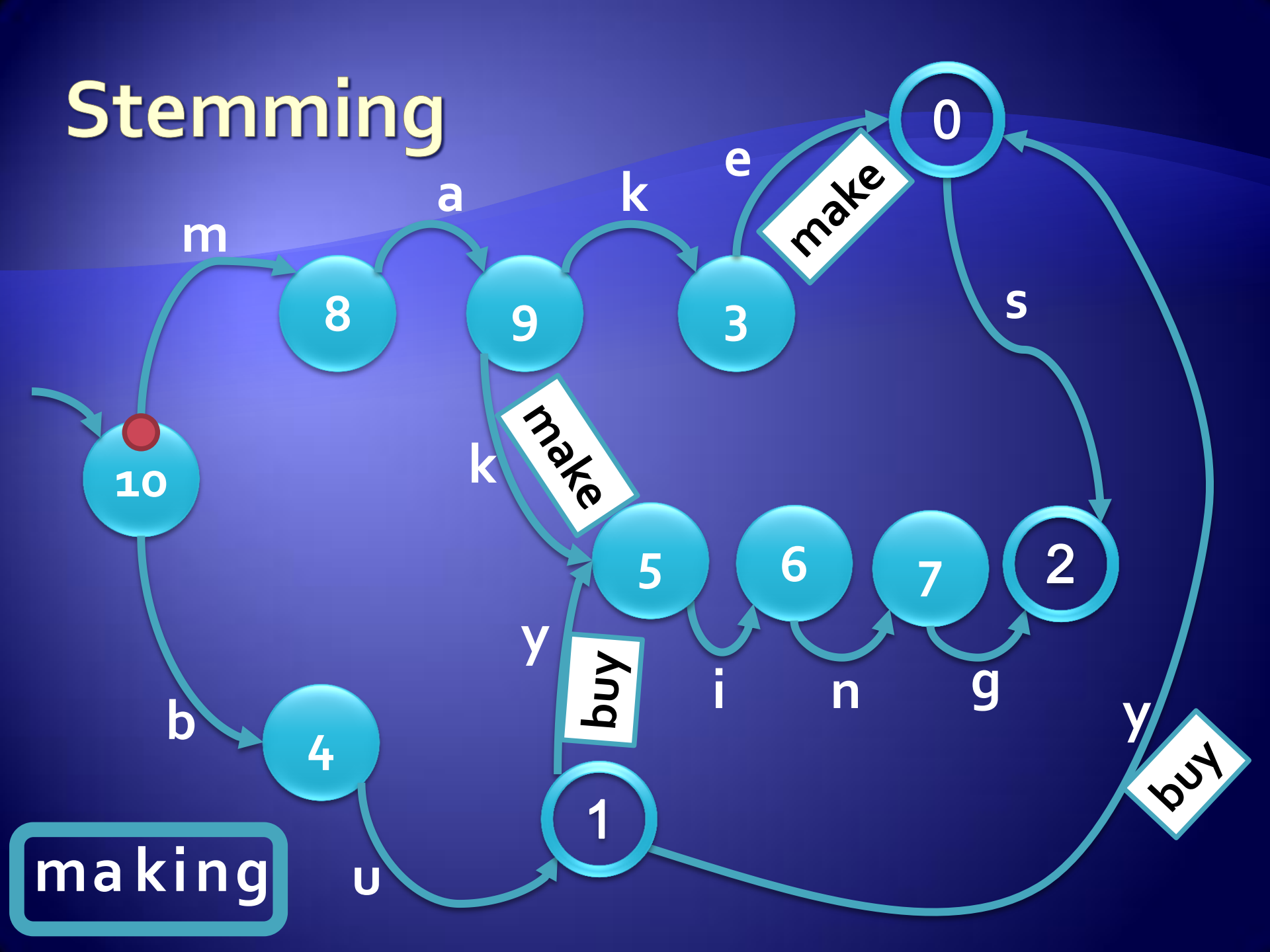

Determine Current State
Check Arcs
Traverse Correct Arc
Set Output String

Output String Storage



Output String Storage

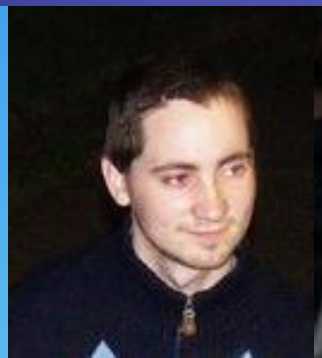
Stemming



Köszönöm!!

Sárközy Gábor
Kornai András
Zsibrita János
Recski Gábor
Erdélyi Miklós

Stanley Selkow
Varga Dániel
Richard Farkas
Zseder Attila
Szabó Adrienne



The SZTAKI Computers