

OPEN SOURCE NATURAL LANGUAGE PROCESSING

Kara Greenfield

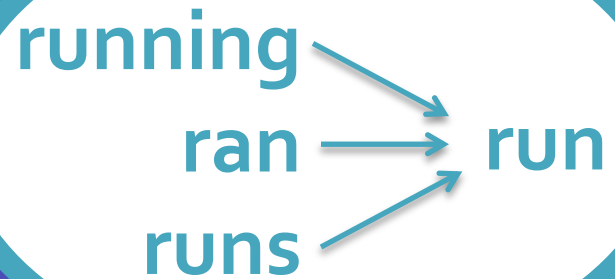
Sarah Judd



Objectives



Finite State Machines
from aff / dic



Stemming

cat ✓
cab ✓
afdd ✗

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

play/GD

make/G

...

making

Flag

Strip

Condition

Complex Morphology

egész

ség

ed

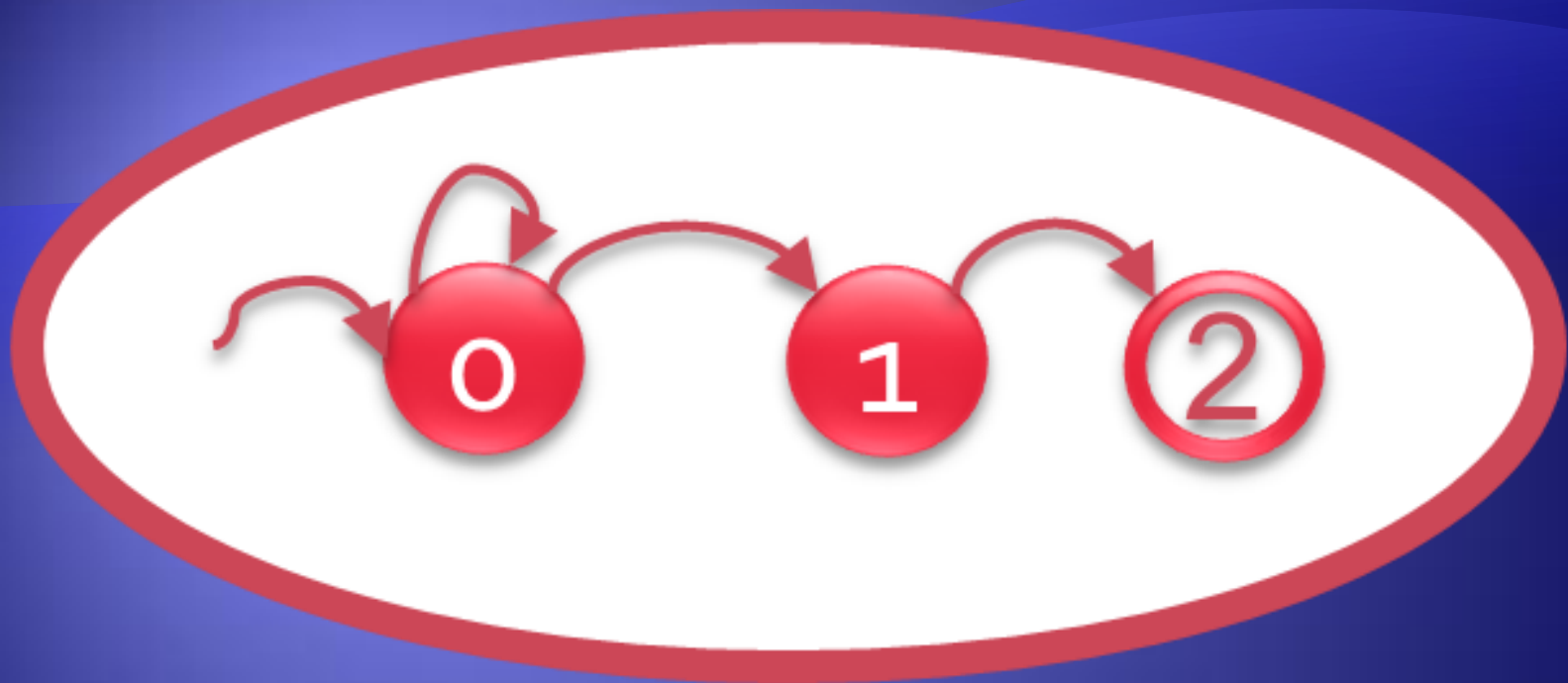
re

egészségedre

to

your

health



Finite State Machines
from aff / dic

Residual Language

$L = \text{English}$
 $u = \text{zo}$

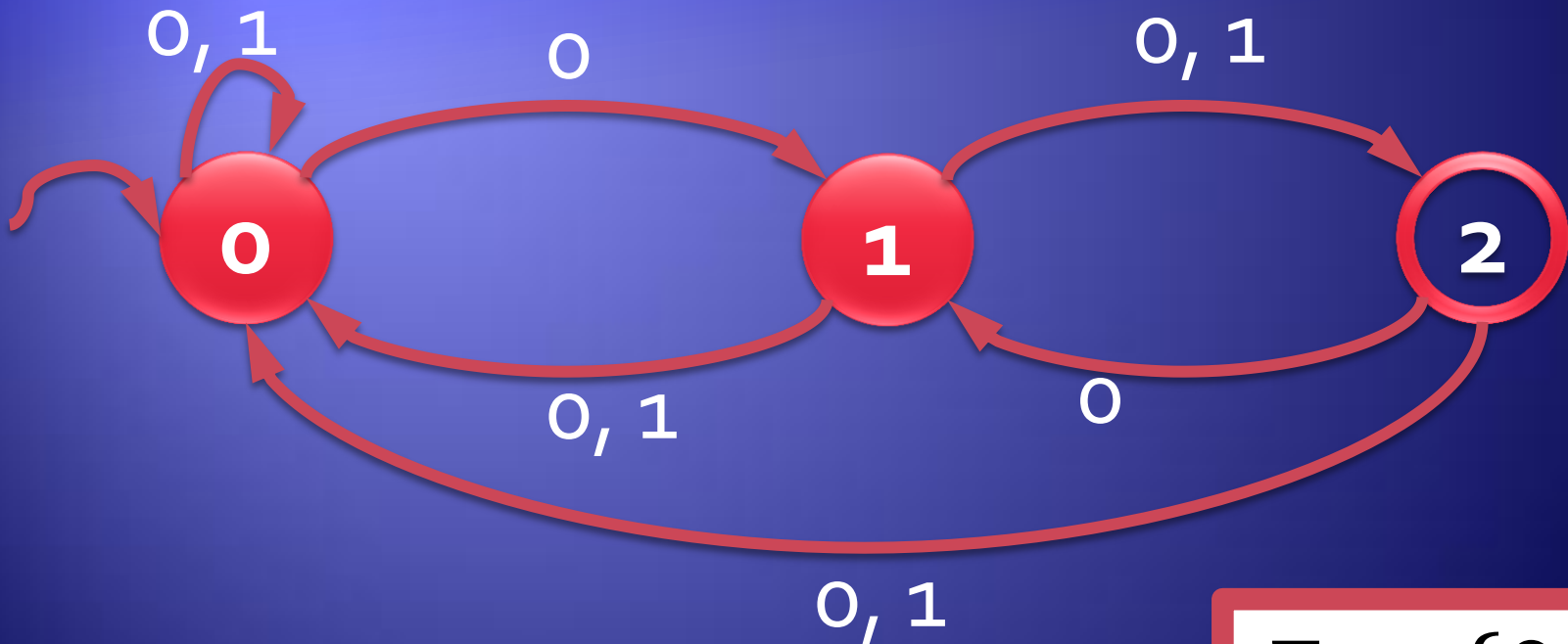
zone
zombie
zodiac
zoo
zoom
...

ne
mbie
diac
o
om

Residual Language

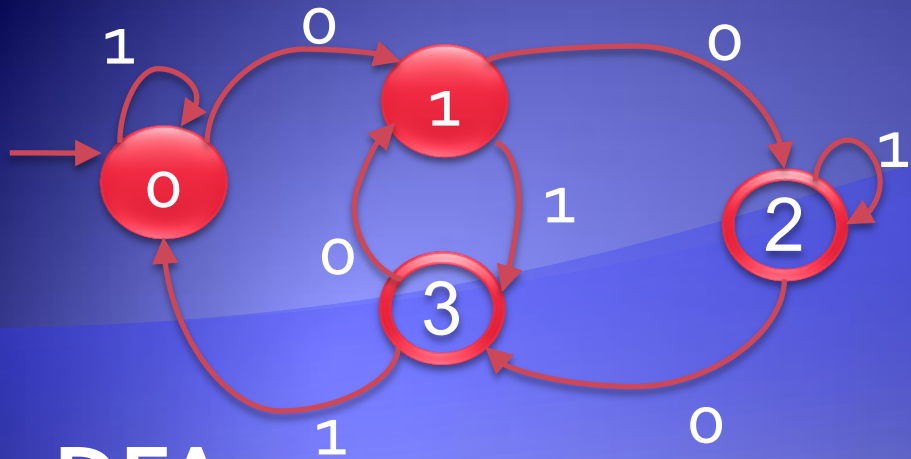
- ◆ Let L be a language over Σ^*
- ◆ Let $u \in \Sigma^*$
- ◆ $u^{-1}L = \{v \in \Sigma^* \mid uv \in L\}$

Residual Finite State Automata



$$\Sigma = \{0, 1\}$$

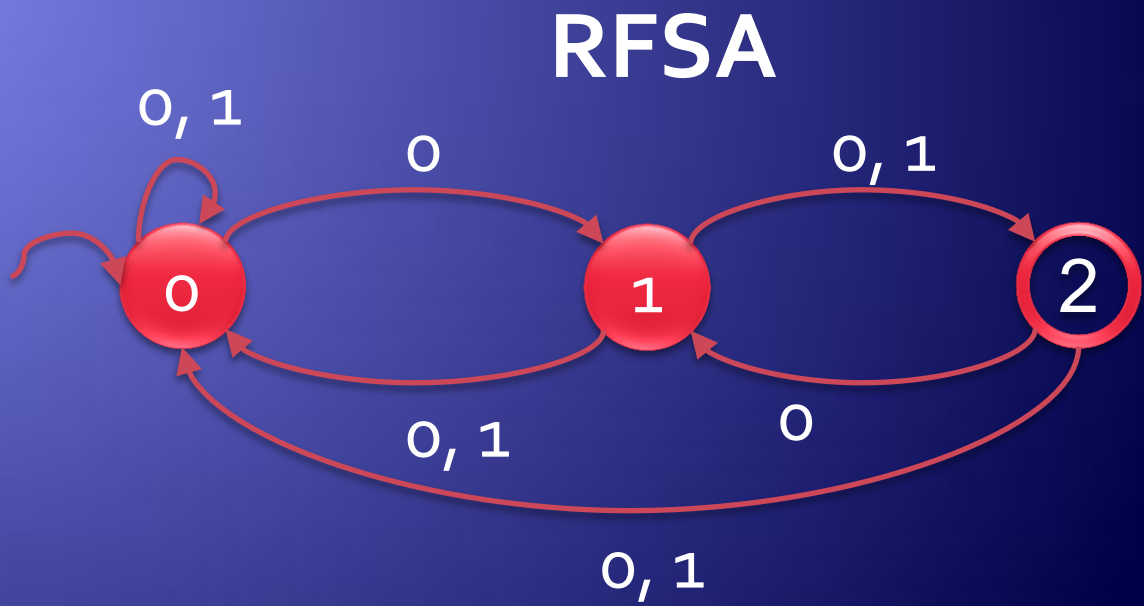
$$L = \Sigma^* 0 \Sigma$$



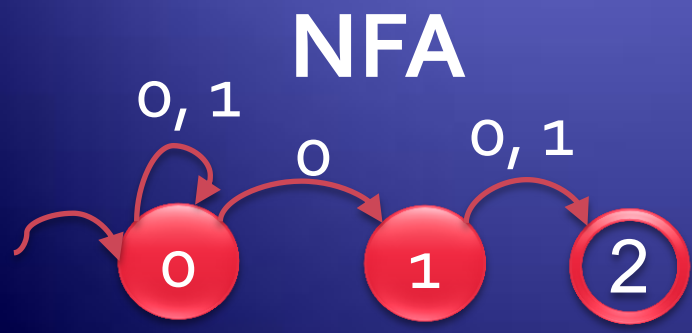
DFA

$$\Sigma = \{0, 1\}$$

$$L = \Sigma^* 0 \Sigma$$



RFSA



NFA

aff/dic → RFSA

aff / dic



NFA
words

DFA

Minimize

Compress

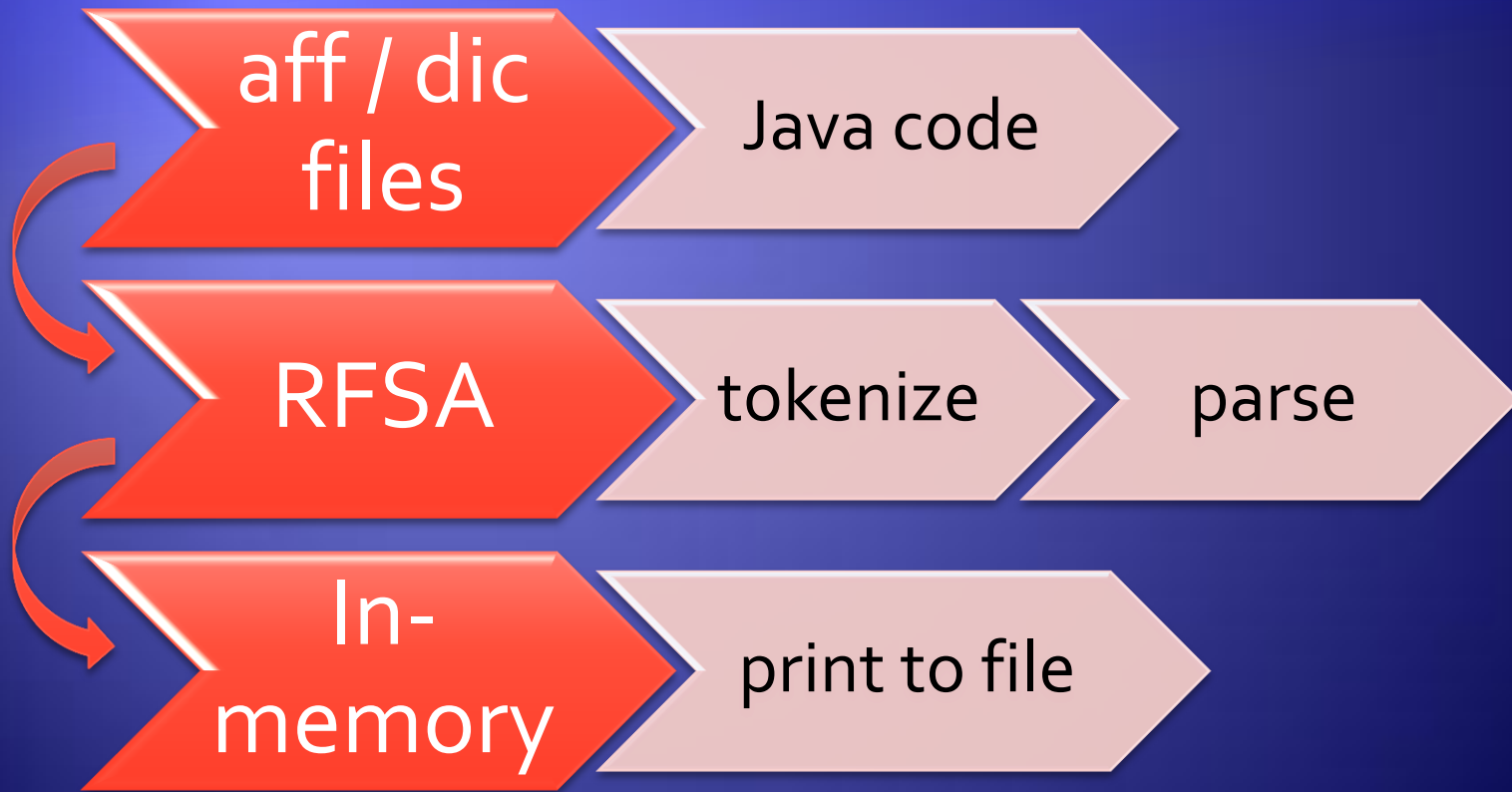
Letterize

Minimize

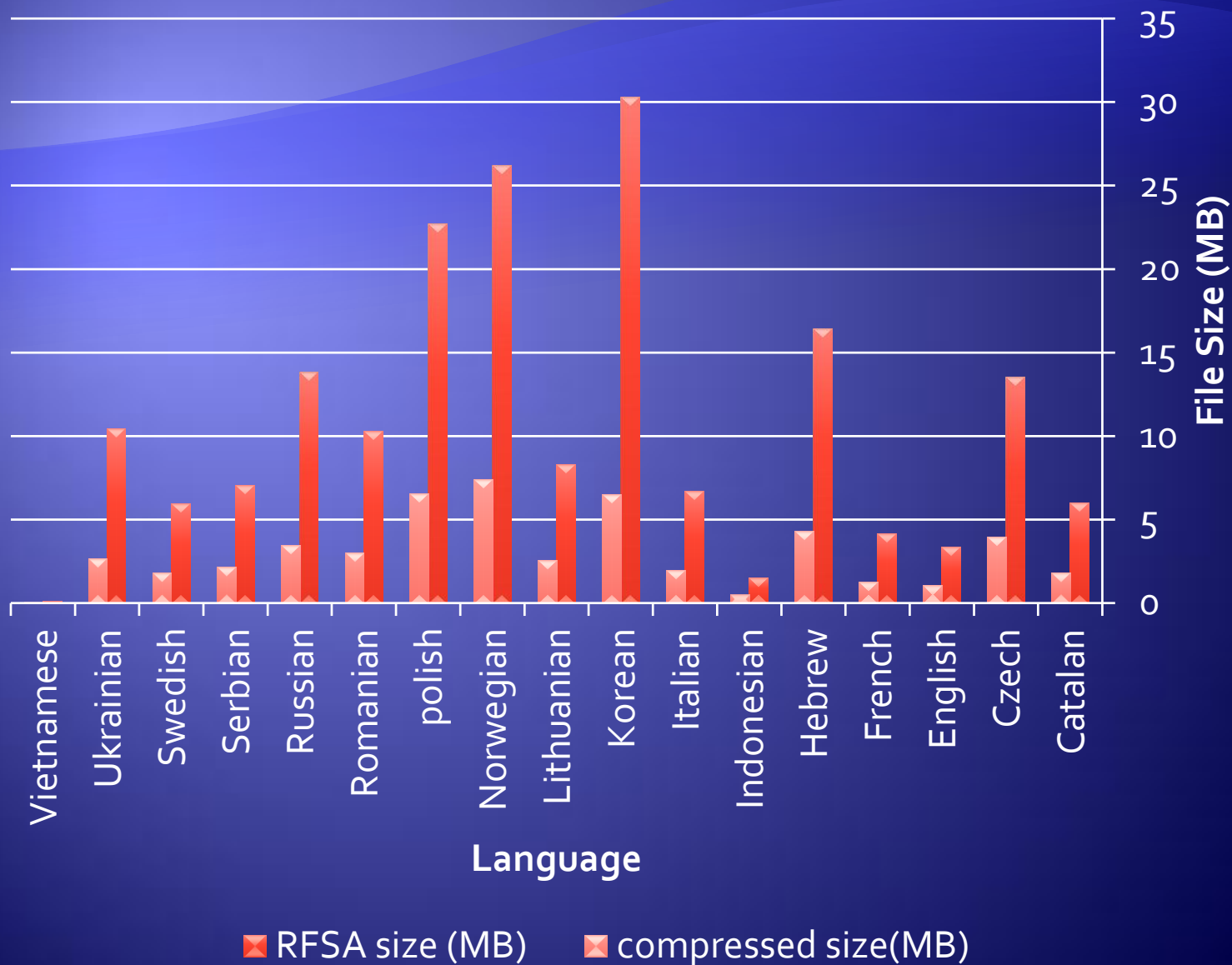
Compress

RFSA

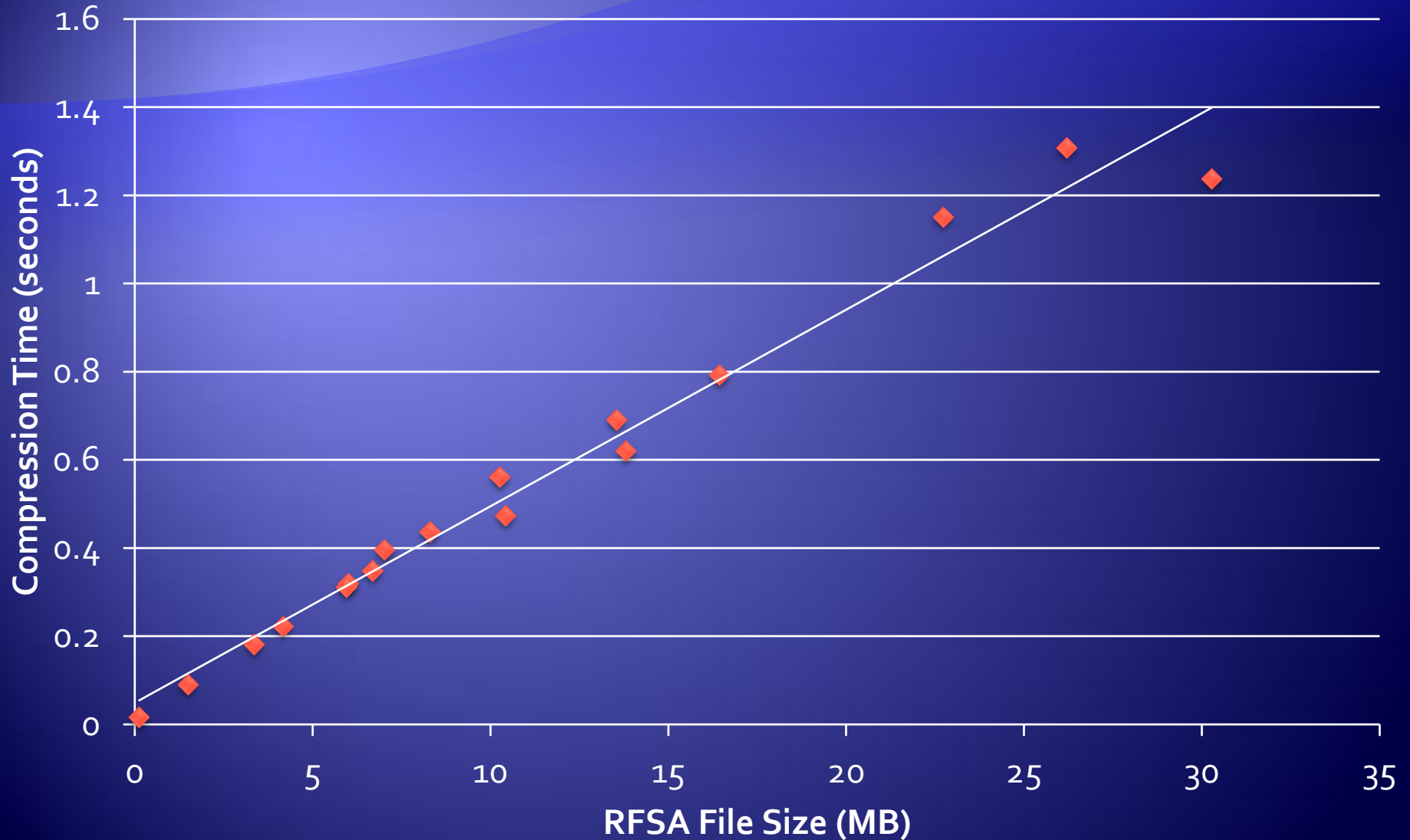
RFSA compression



Language Compression



Language Compression



cat ✓

cab ✓

afdd ✗

Spell Checking

Spell Checking Process

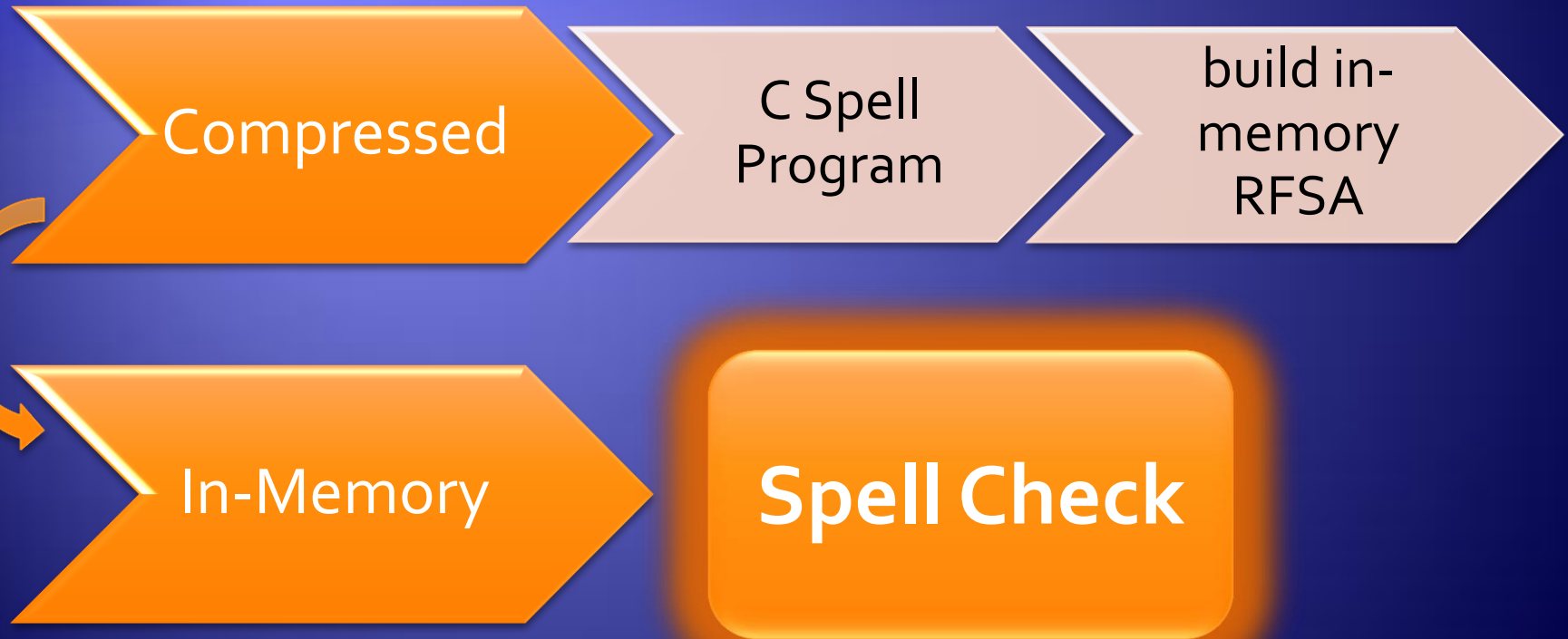
Compressed

C Spell
Program

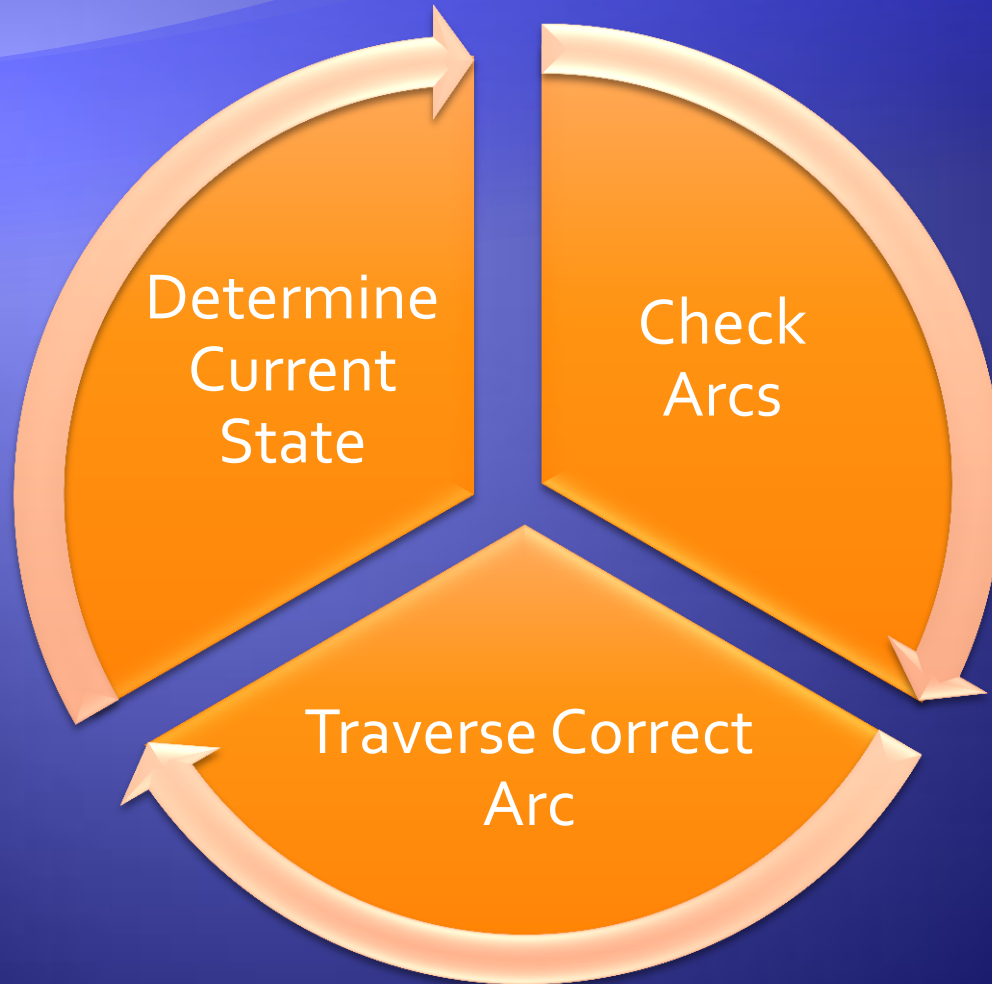
build in-
memory
RFSA

In-Memory

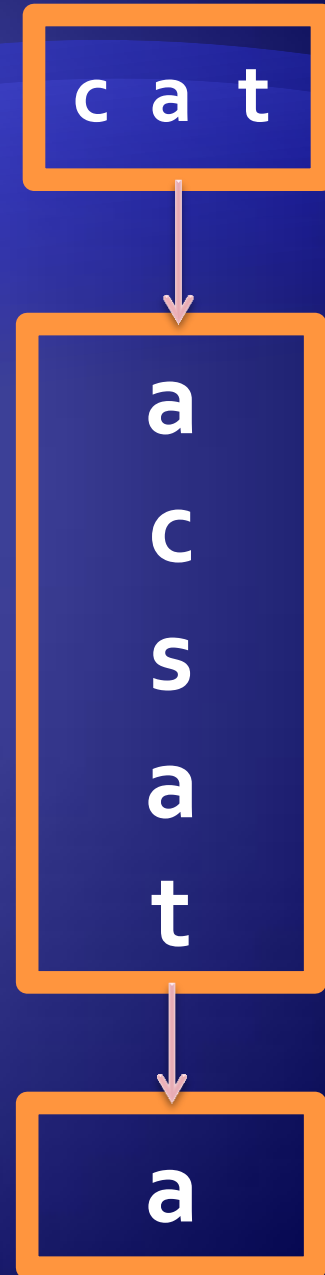
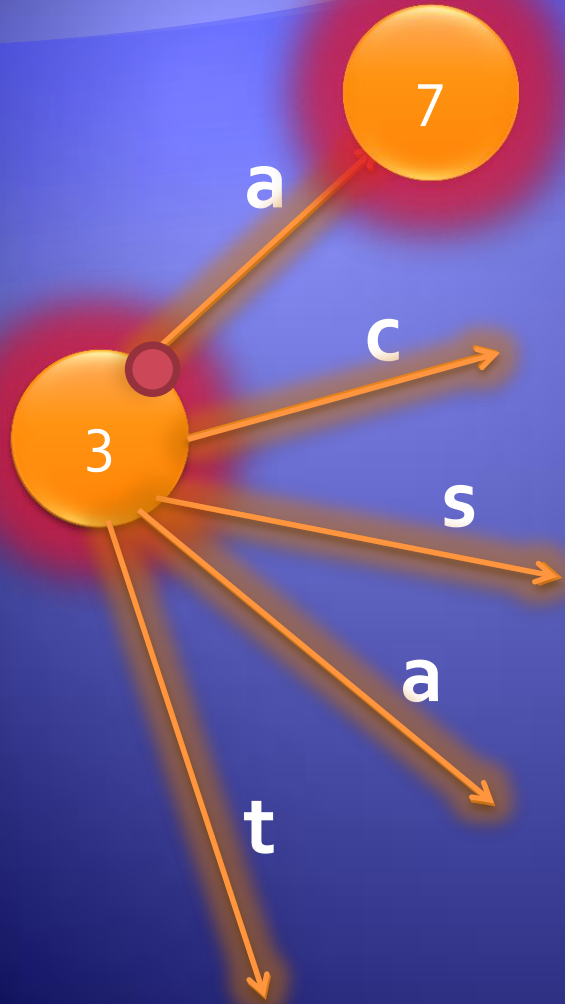
Spell Check



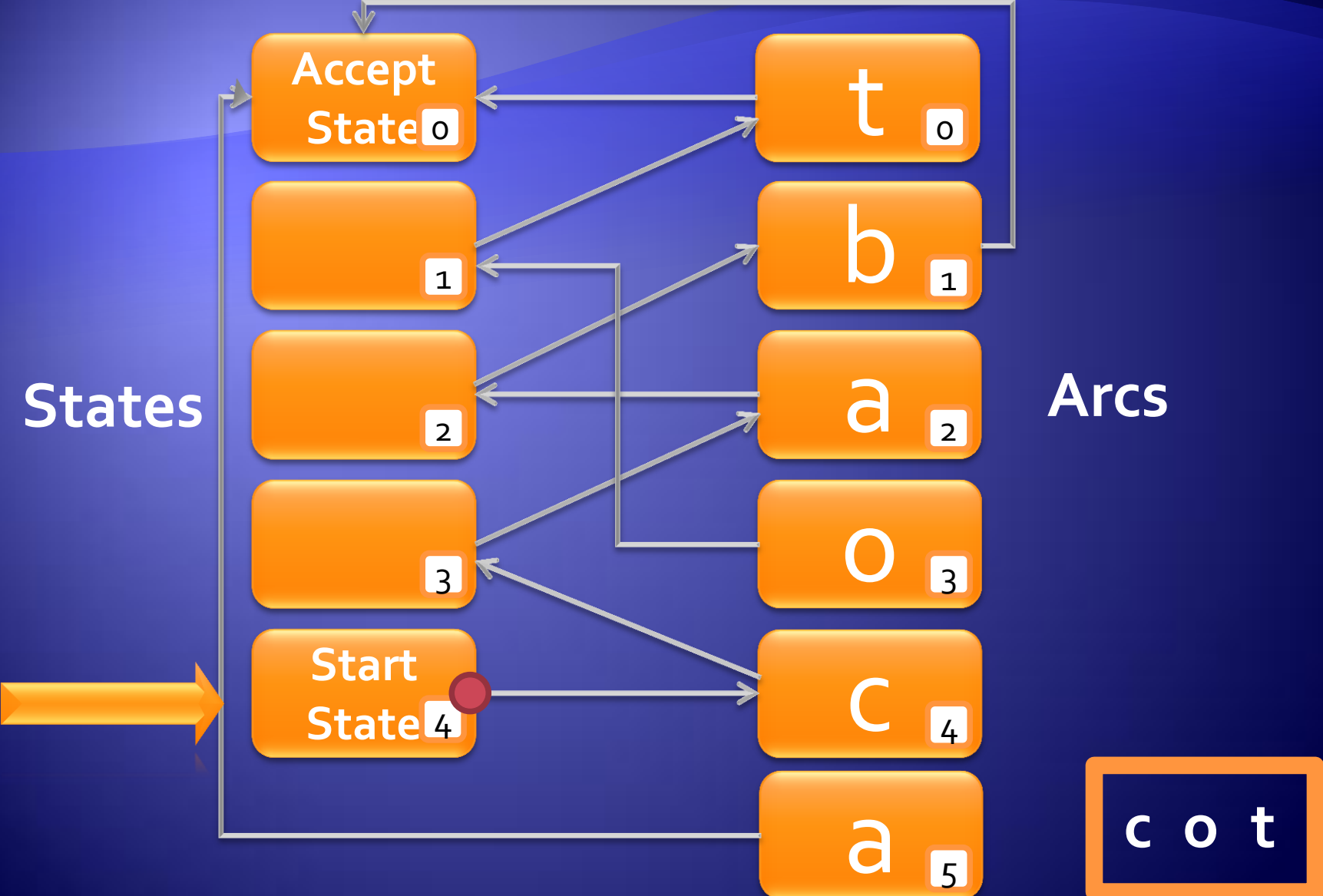
Spell Checking



Spell Checking



Spell Checking



Spell Checking Results



92%



non-English



non-standard abbreviations



spelling errors



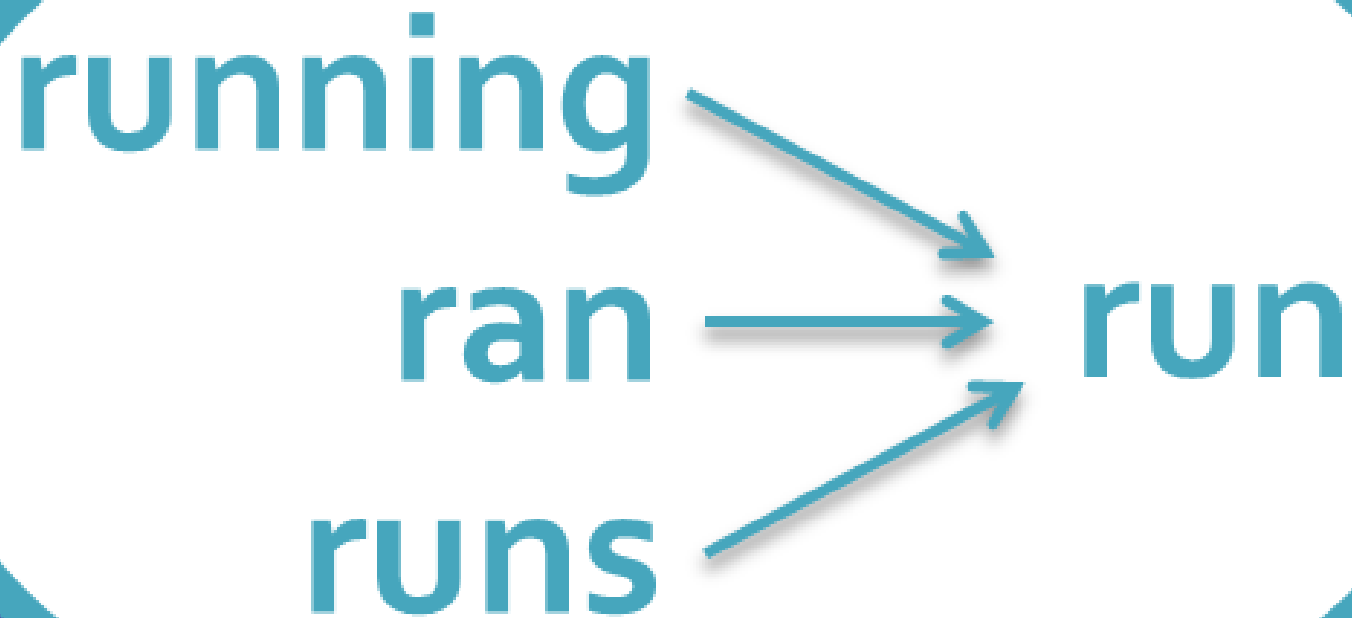
87%



uncommon proper nouns

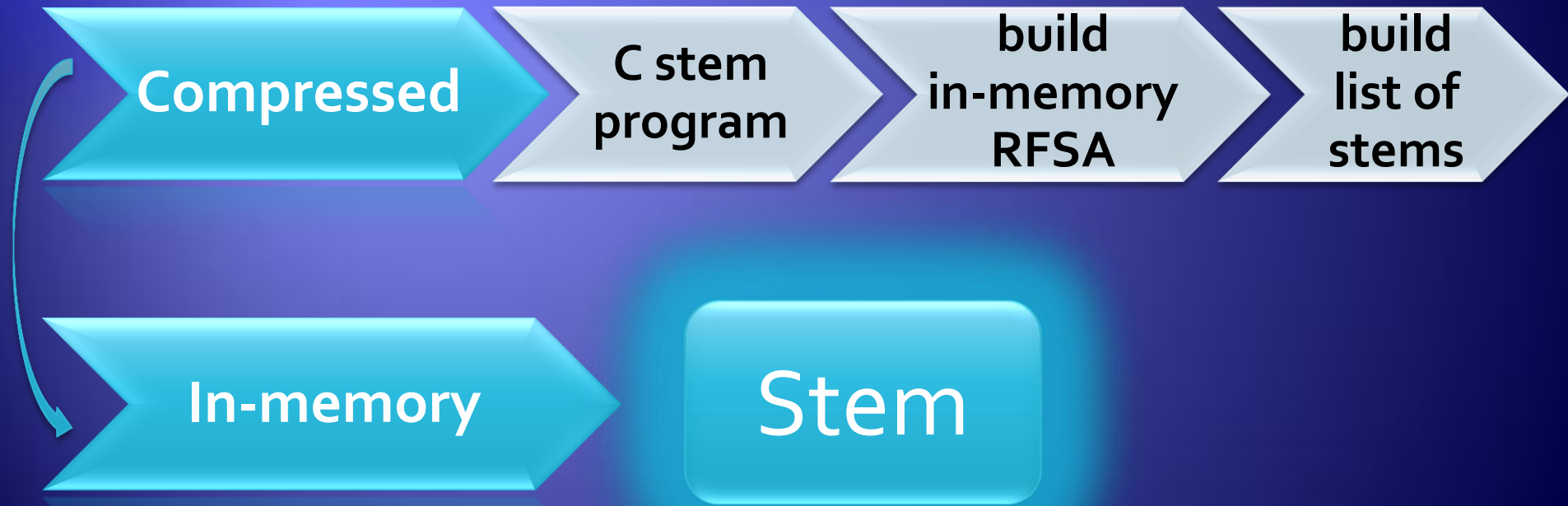


missing from aff/dic

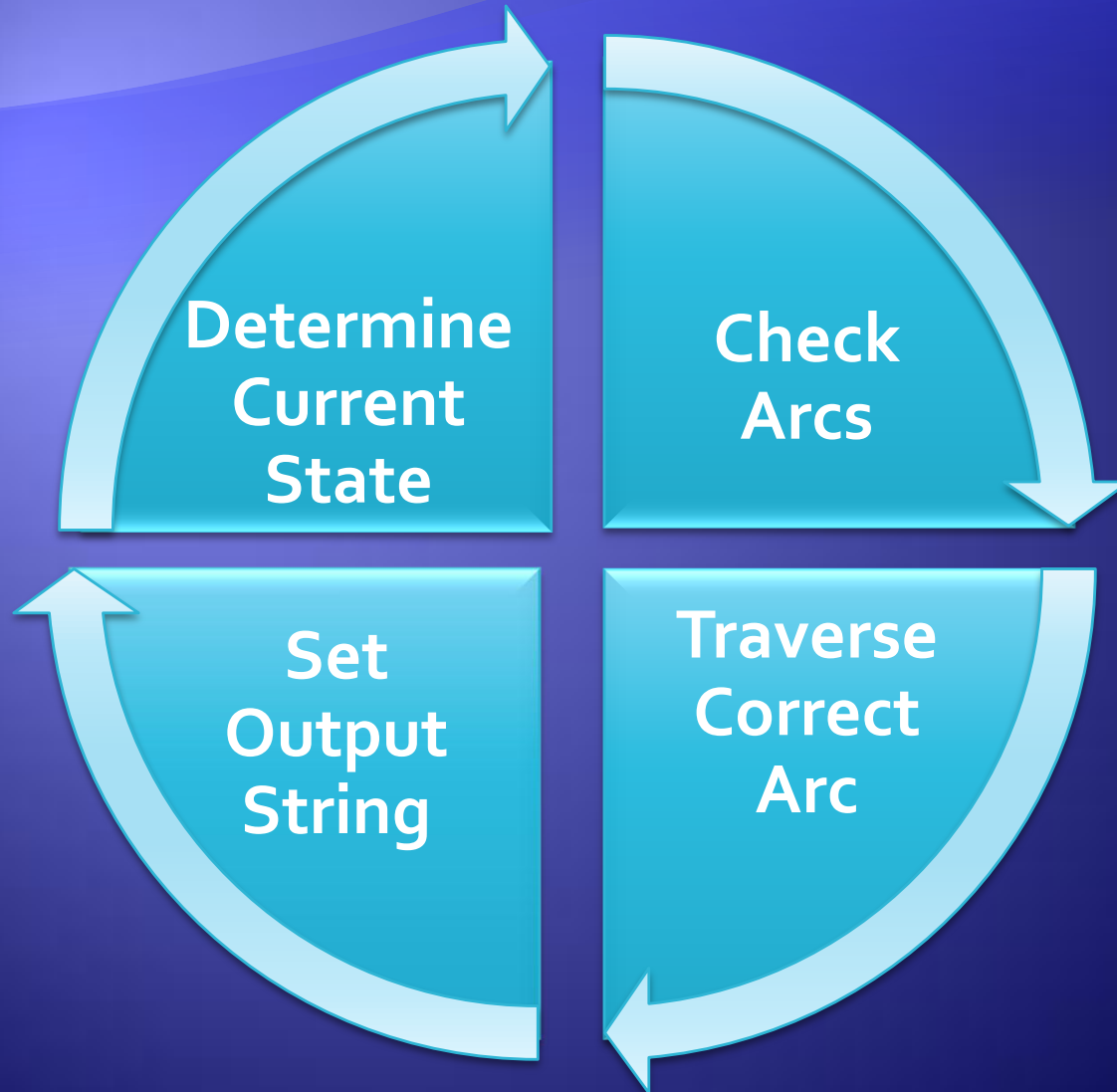


Stemming

Stemming Process



Stemming



Output String Storage



Output String Storage

\$a

\$b

\$g

\$n

\$m

\$y

\$z

\$aa

rdvark

\$ab

le

sent

solute

\$ca

b

r

re

se

t

\$mo

on

re

uth

ve

\$zo

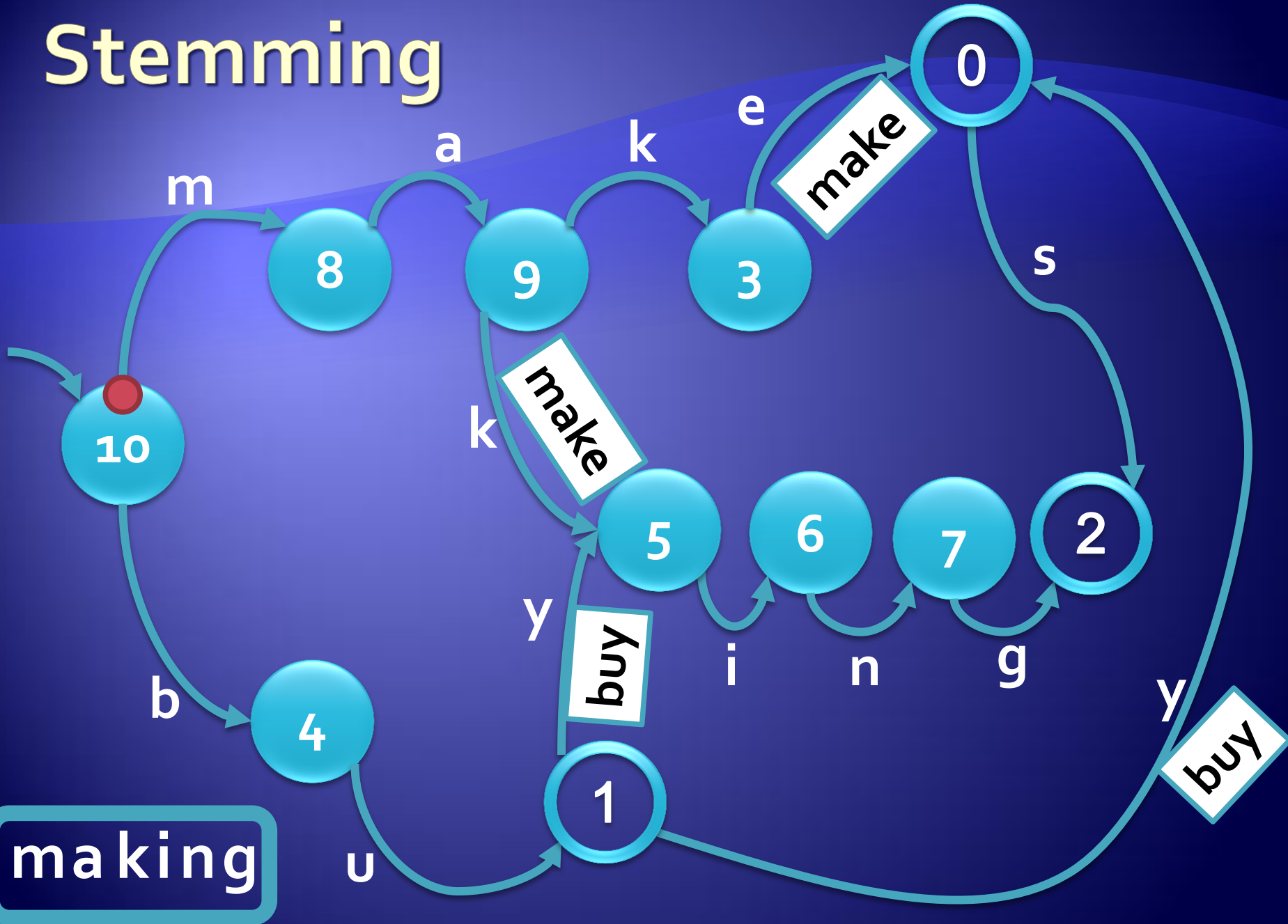
diac

mbie

ne

o

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

```
kgreenfield@nessi5: ~  
top - 11:46:26 up 11 days, 21:31, 14 users,  load average: 4.10, 3.86, 3.23  
Tasks: 252 total,  6 running, 242 sleeping,   0 stopped,   4 zombie  
Cpu(s): 37.8%us,  4.7%sy, 12.4%ni, 35.8%id,  9.0%wa,  0.1%hi,  0.1%si,  0.0%st  
Mem:  99209620k total, 98734708k used,  474912k free,   75524k buffers  
Swap: 16803576k total,  135140k used, 16668436k free,  3090192k cached
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
11722	kgreenfi	20	0	88.6g	88g	2824	R	100	93.5	600:30.28	emacs
16700	kgreenfi	20	0	83360	2700	1728	S	0	0.0	0:00.00	sshd
16701	kgreenfi	20	0	49972	2076	2072	S	0	0.0	0:00.00	sftp-server
18641	kgreenfi	20	0	81176	2500	1524	S	0	0.0	0:00.01	ssh
18642	kgreenfi	20	0	40736	5744	2160	S	0	0.0	0:00.14	bash
19034	kgreenfi	20	0	7836	2384	1672	R	0	0.0	0:00.49	top

88g

100%

emacs