Smaller RLZ-Compressed Suffix Arrays

Simon J. Puglisi and Bella Zhukova

DCC 2021

Suffix Sorting

Suffix Sorting

0	1	2	3	4	5	6
a	b	a	a	b	a	b

Suffix Sorting

0	1	2	3	4	5	6
a	b	a	a	b	a	b

0	a b a a b a b
1	b a a b a b
2	a a b a b
3	a b a b
4	b a b
5	a b
6	b

Suffix Sorting

0	1	2	3	4	5	6
a	b	a	a	b	a	b

0	a b a a b a b	2	a a b a b
1	b a a b a b	5	a b
2	a a b a b	0	a b a a b a b
3	a b a b	3	a b a b
4	b a b	6	b
5	a b	1	b a a b a b
6	b	4	b a b

Suffix Sorting

0	1	2	3	4	5	6
a	b	a	a	b	a	b

0	a b a a b a b	2	a a b a b
1	b a a b a b	5	a b
2	a a b a b	0	a b a a b a b
3	a b a b	3	a b a b
4	b a b	6	b
5	a b	1	b a a b a b
6	b	4	b a b

Problem (Pattern Matching)

Find all occurrences of a pattern P in a text T

Popular solution: find the interval of the suffix array (SA) that contains them

- Binary search using SA and text, or
- Backward search on the Burrows-Wheeler Transform of T (FM-index)
- Lots of compressed versions of the SA
  - Problem then becomes: how do we decompress the interval's contents?

Previous Work

- LCSA (González, Navarro, Ferrada, Journal of Experimental Algorithmics 2014)
  - uses similar ideas as rlzsa but with RePair grammar compression, wasn't compared to r-index before
- r-index (Gagie et al., SODA 2018)
  - recent, very fast, very small — a huge leap forward in compressed indexing
- rlzsa: RLZ compression (Puglisi and Zhukova, SPIRE 2020)
  - significantly faster than r-index, takes more space

Our Contribution

- Better reference selection
  - in previous work good results with randomly generated references
  - here — a deterministic method based on k-mer frequencies
- Compact representation of index components

Core Idea

SA					30	25	20	15	10	5									29	24	19	14	9							

repetitions that are off by 1 (Mäkinen, CPM 2000)

differences will turn into actual repetitions (González, Navarro, CPM 2007)

Overview of the Algorithm

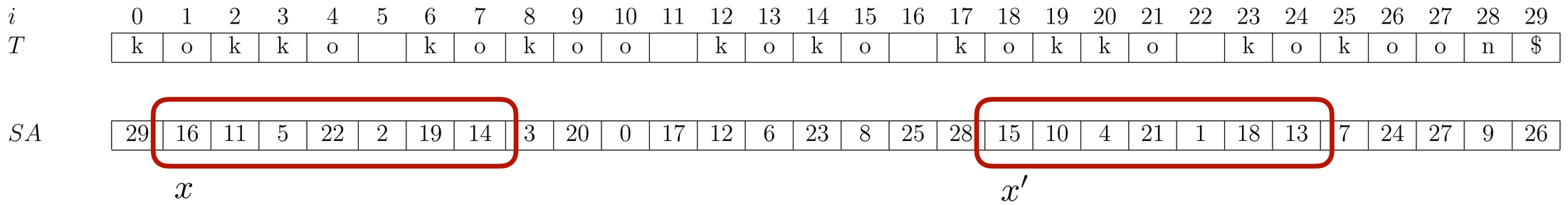
Compression:

1. form differentially encoded SA^{diff} from SA
2. form reference R by selecting substrings from SA^{diff}
3. use Relative Lempel-Ziv (RLZ) to parse SA^{diff} relative to R
4. output reference R plus set of phrases (pointers into R)

Decompression requires:

1. predecessor data structure containing phrase starting positions (in order to find the phrase covering the start of an interval)
2. absolute SA value for a starting position of the phrase

Example



$SA[x, y]$ preceded by symbol $c \Rightarrow$

$\exists SA[x', x' + (y - x)]$:

$\forall i \in [0, y - x] \quad SA[x + i] = SA[x' + i] + 1$

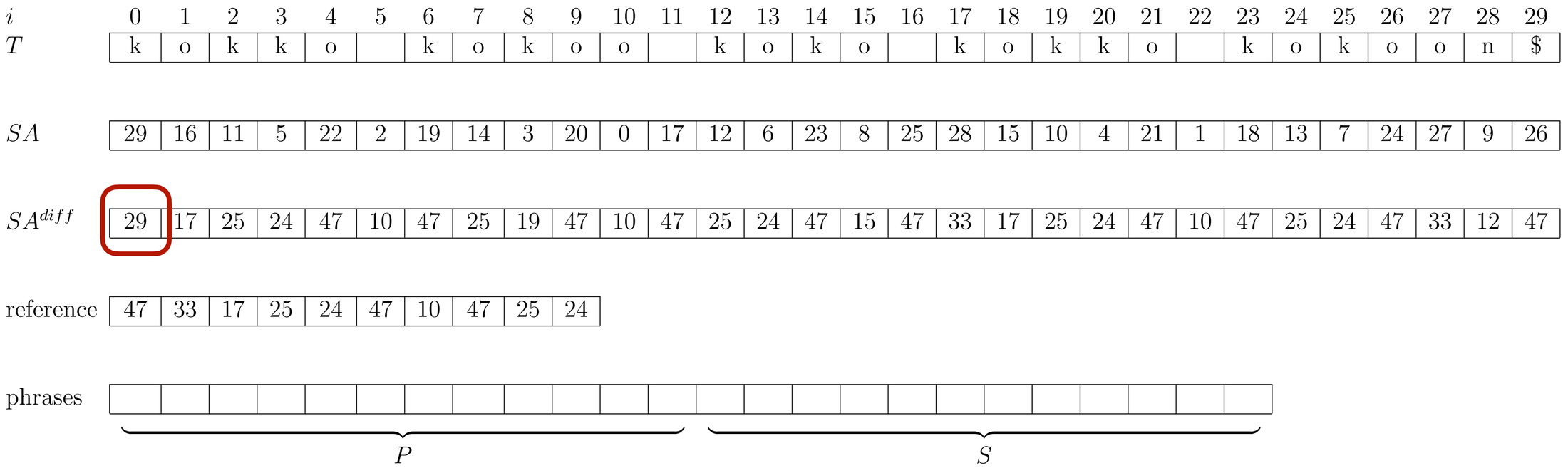
(González and Navarro, CPM 2007)

Example

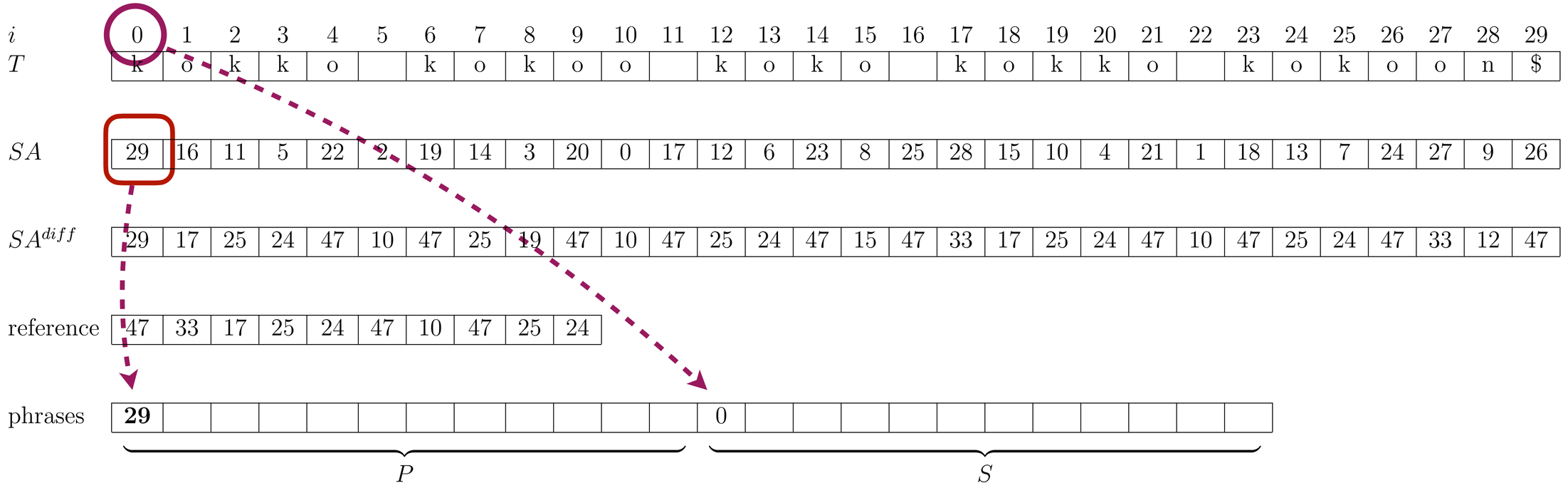
i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
T	k	o	k	k	o		k	o	k	o	o		k	o	k	o		k	o	k	k	o		k	o	k	o	o	n	$
SA	29	16	11	5	22	2	19	14	3	20	0	17	12	6	23	8	25	28	15	10	4	21	1	18	13	7	24	27	9	26
SA^{diff}	29	17	25	24	47	10	47	25	19	47	10	47	25	24	47	15	47	33	17	25	24	47	10	47	25	24	47	33	12	47

$$i \in [1, n-1] \qquad SA^{diff}[i] = SA[i] - SA[i-1] + n$$

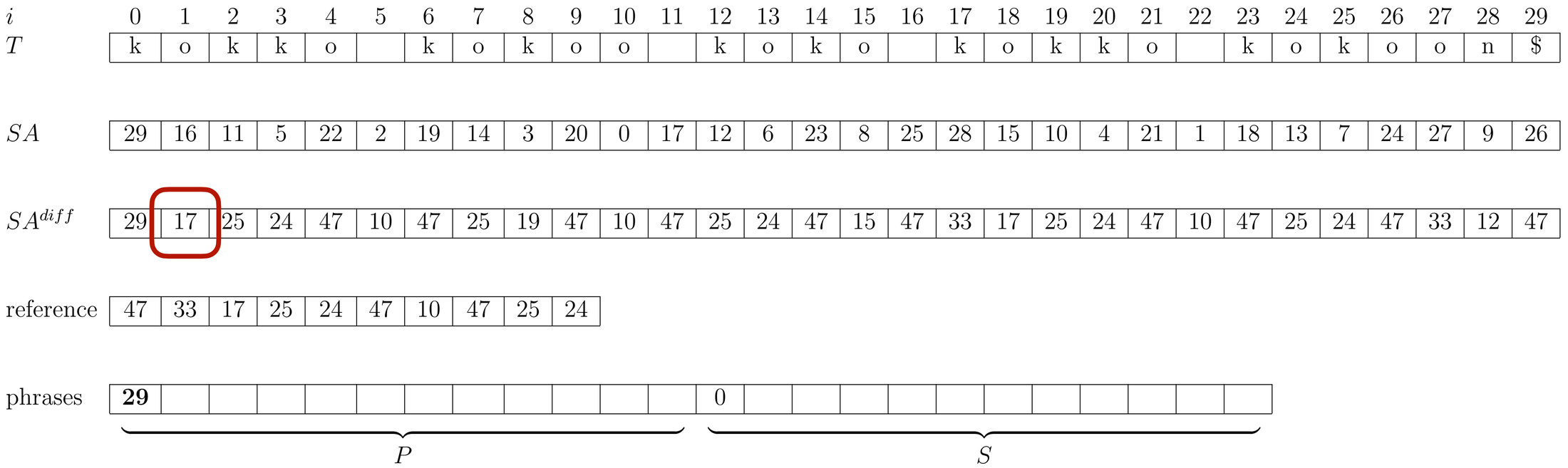
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases
P
S

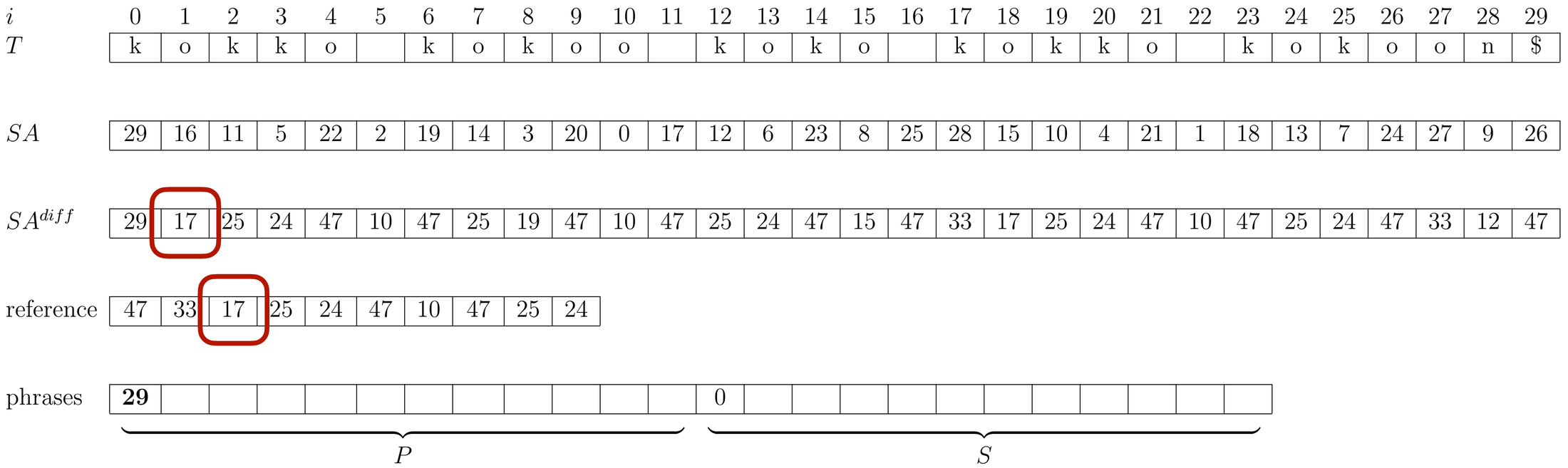
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k k o k o k k o k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 0
P
S

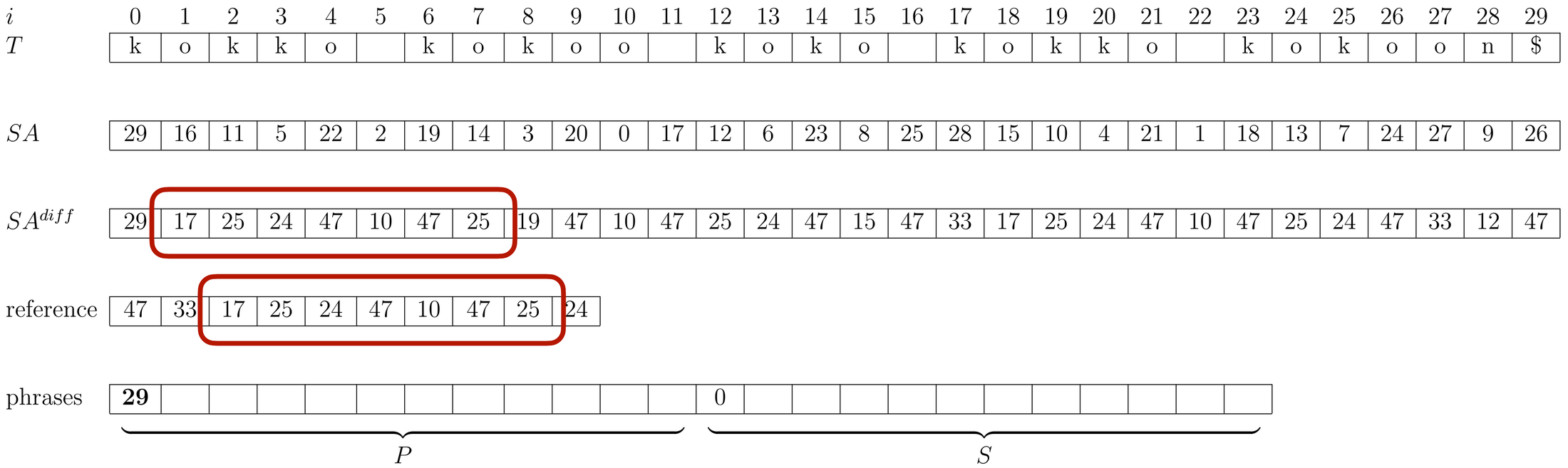
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 0
P
S

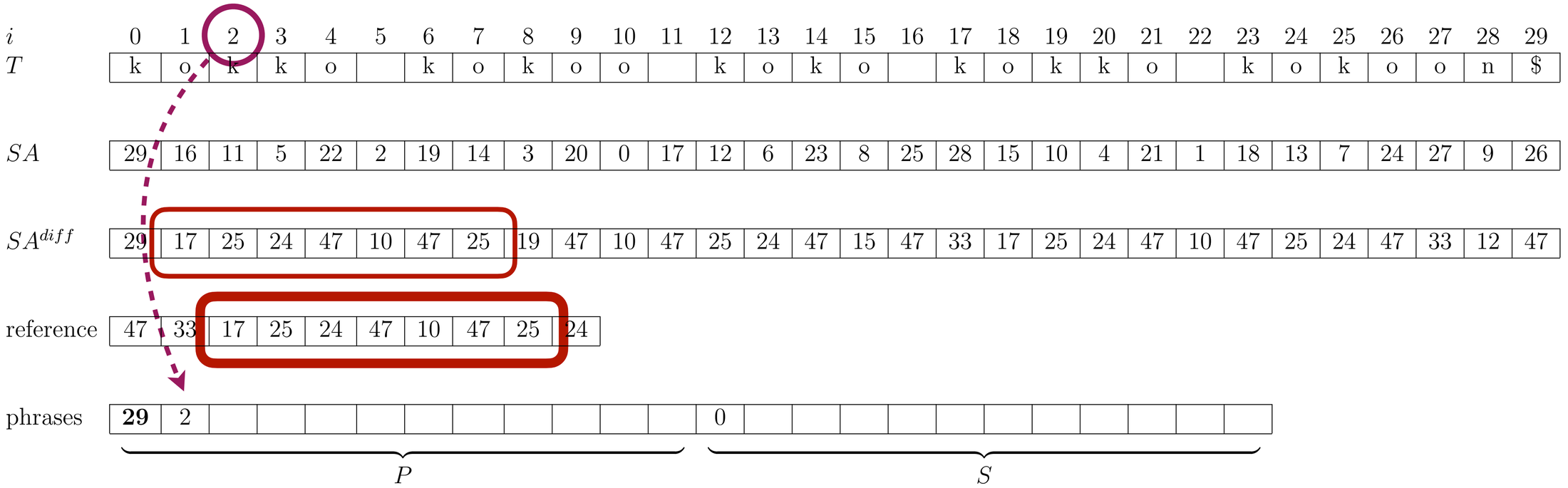
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 0
P
S

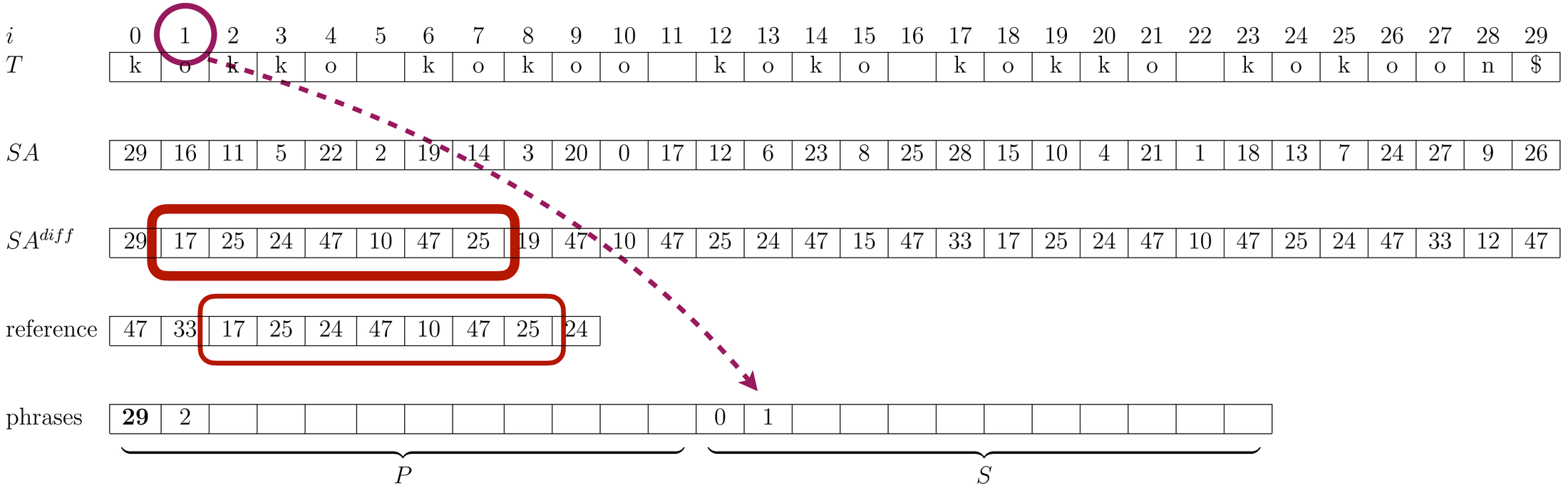
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 0
P
S

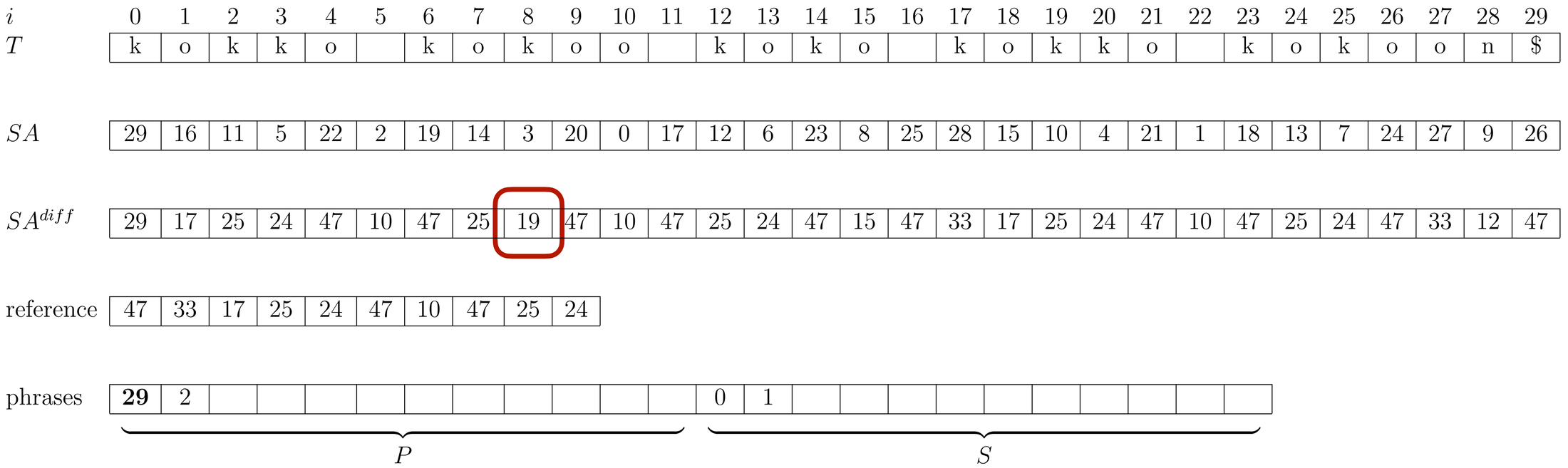
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k k o k o k k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 0
P
S

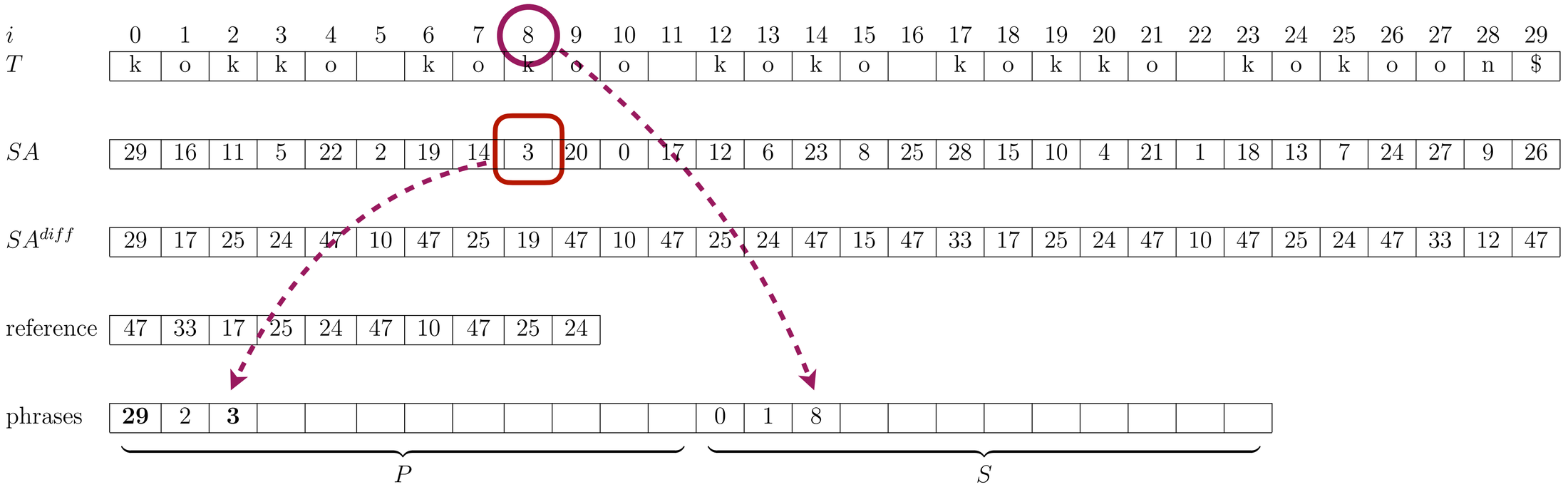
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k k o k o k k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 0 1
P
S

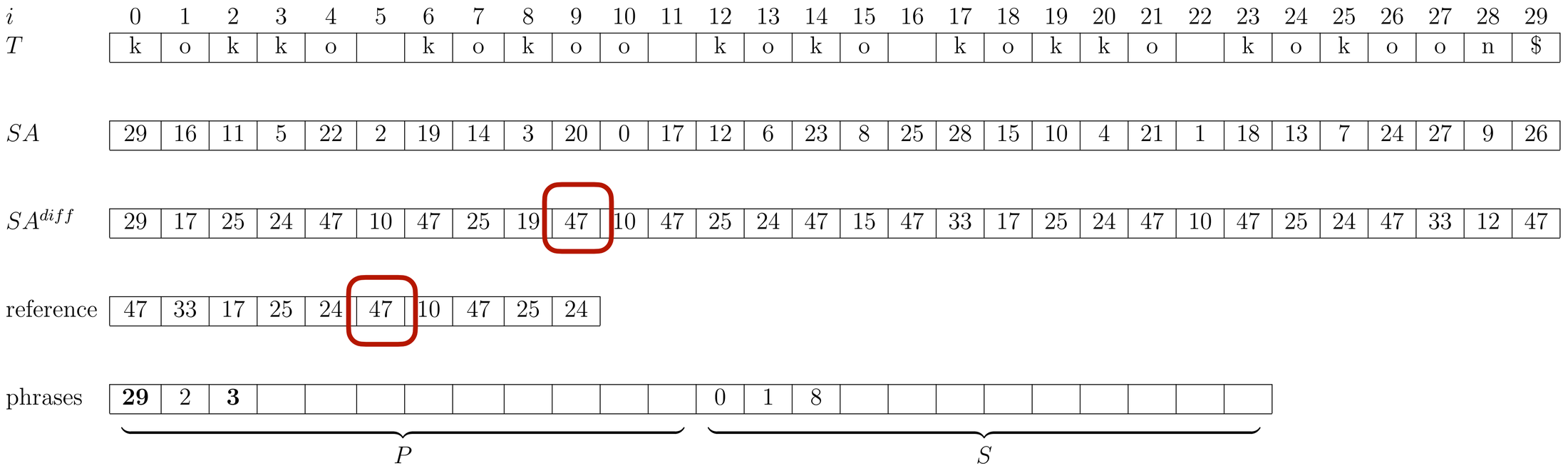
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 0 1
P
S

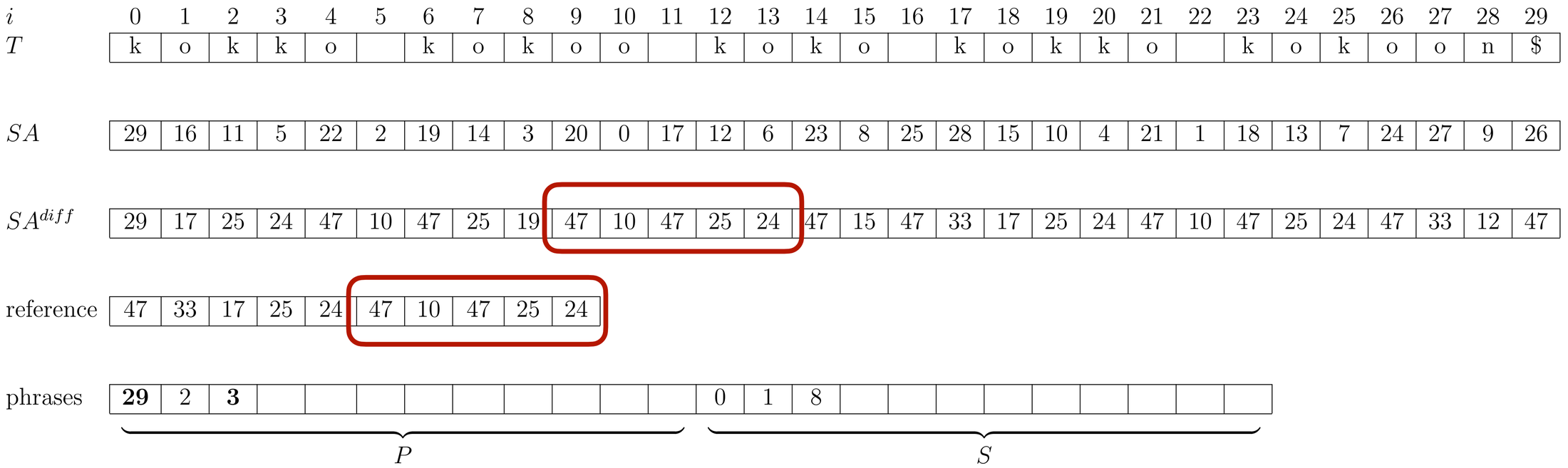
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k k o k k o $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 0 1 8
P
S

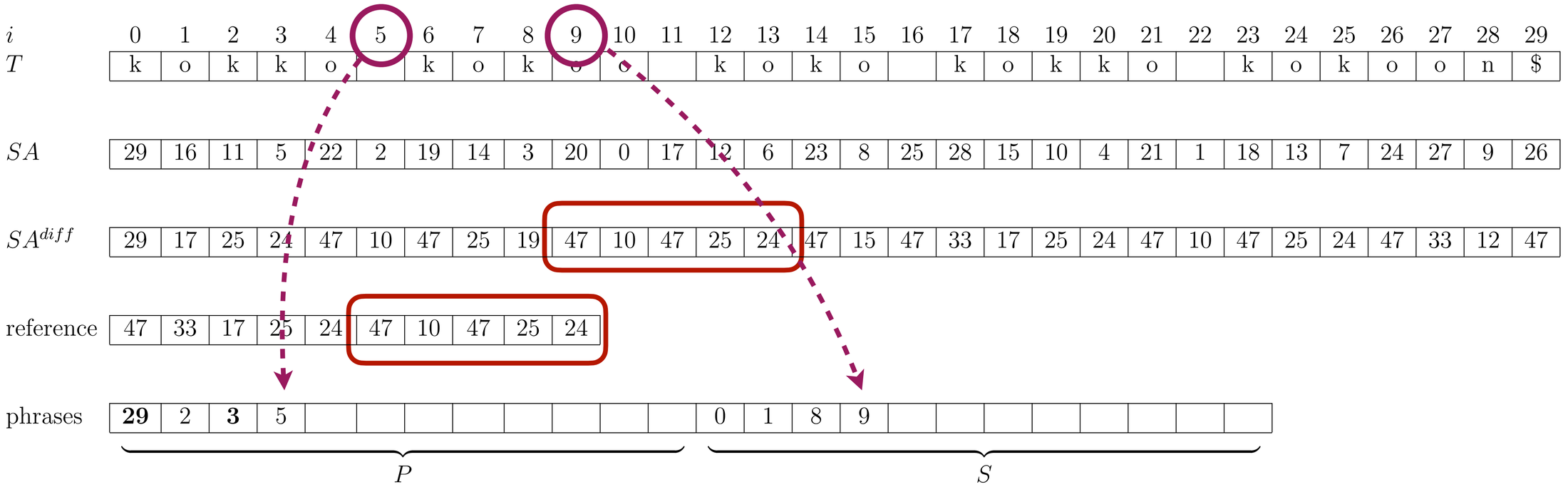
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 0 1 8
P
S

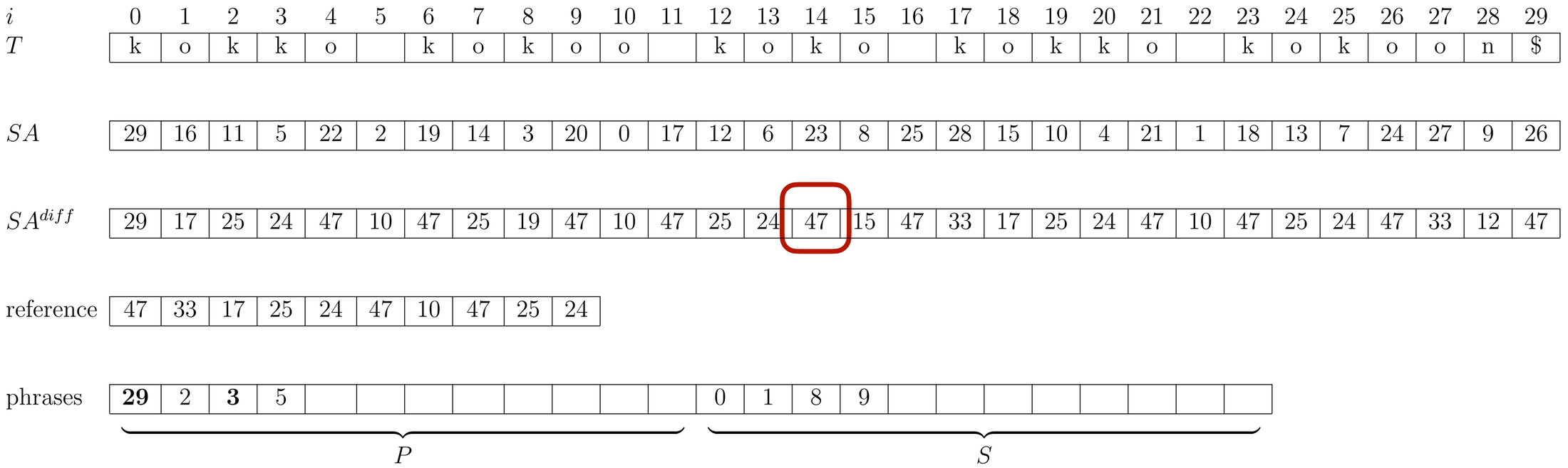
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k k o k k o k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 0 1 8
P
S

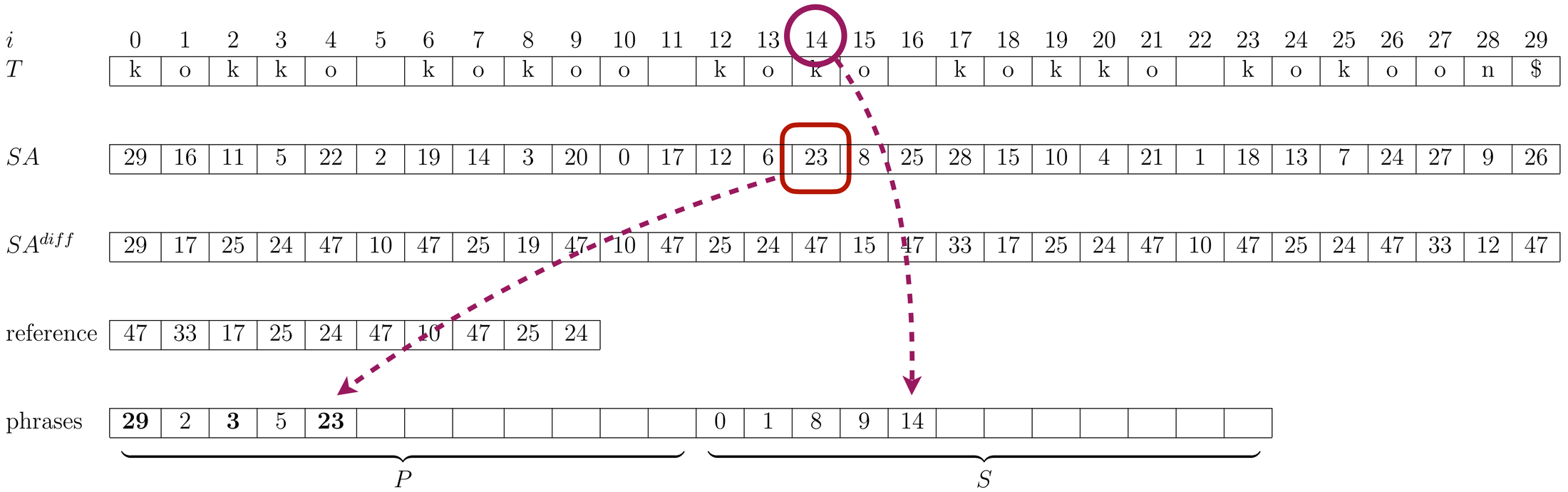
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 0 1 8 9
P
S

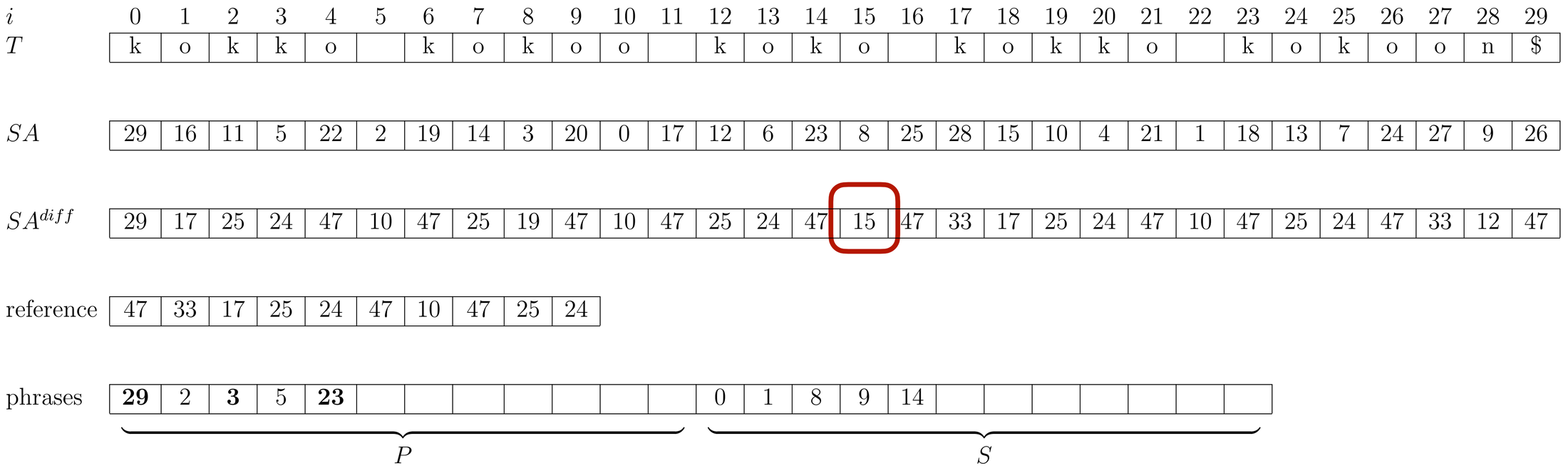
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 0 1 8 9
P
S

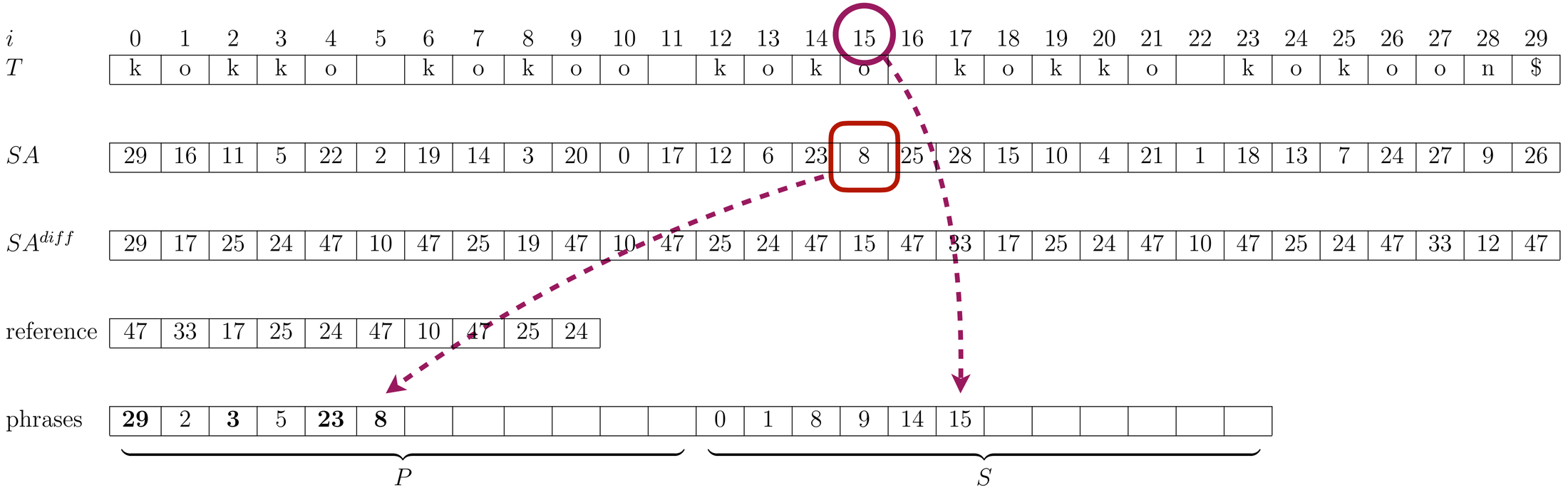
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 0 1 8 9 14
P
S

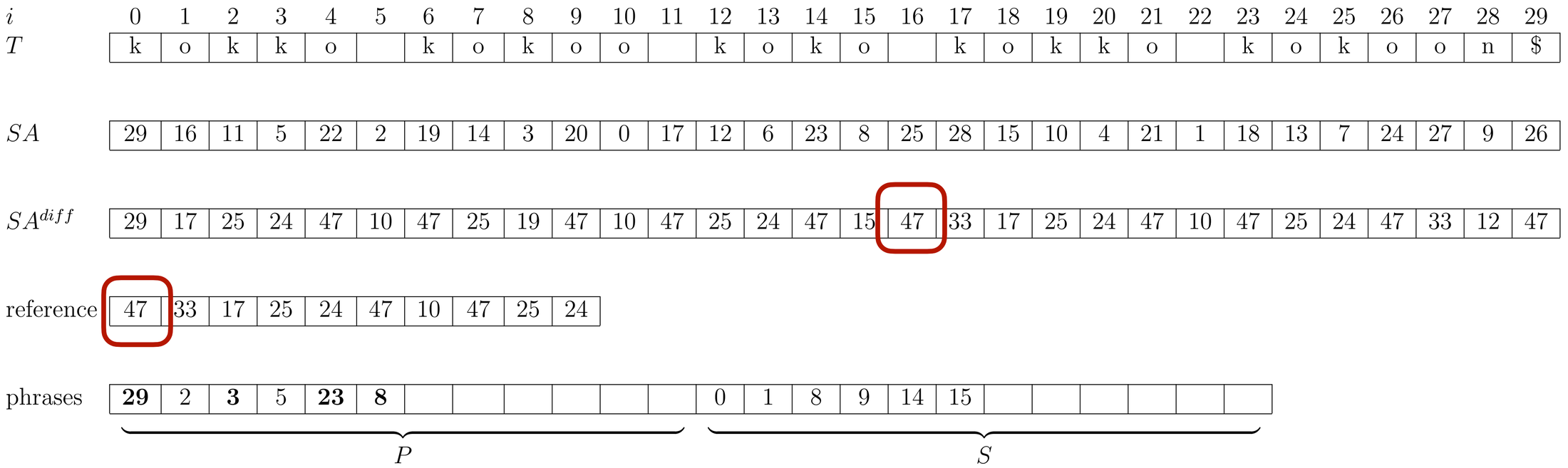
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 0 1 8 9 14
P
S

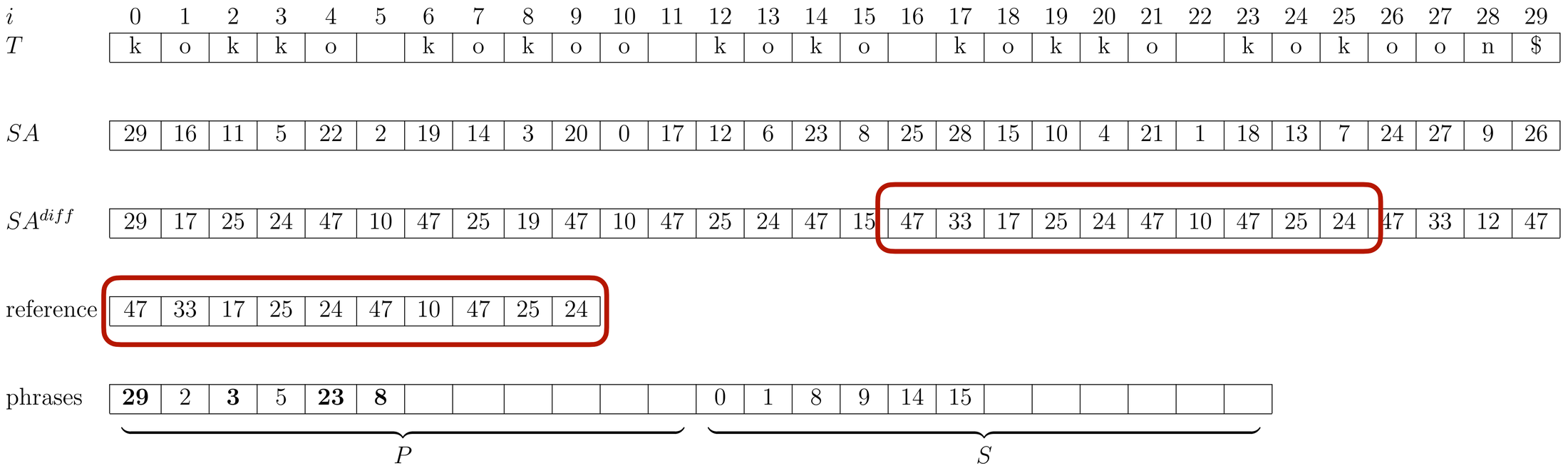
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 1 8 9 14 15
P
S

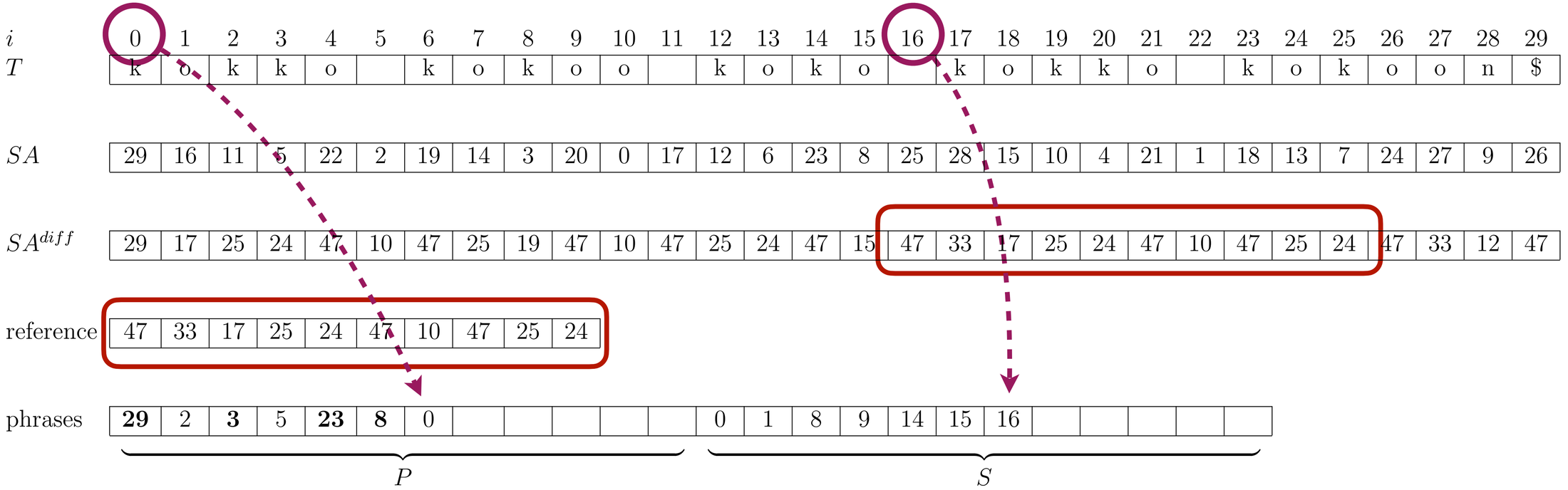
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k k o k o k k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 1 8 9 14 15
P
S

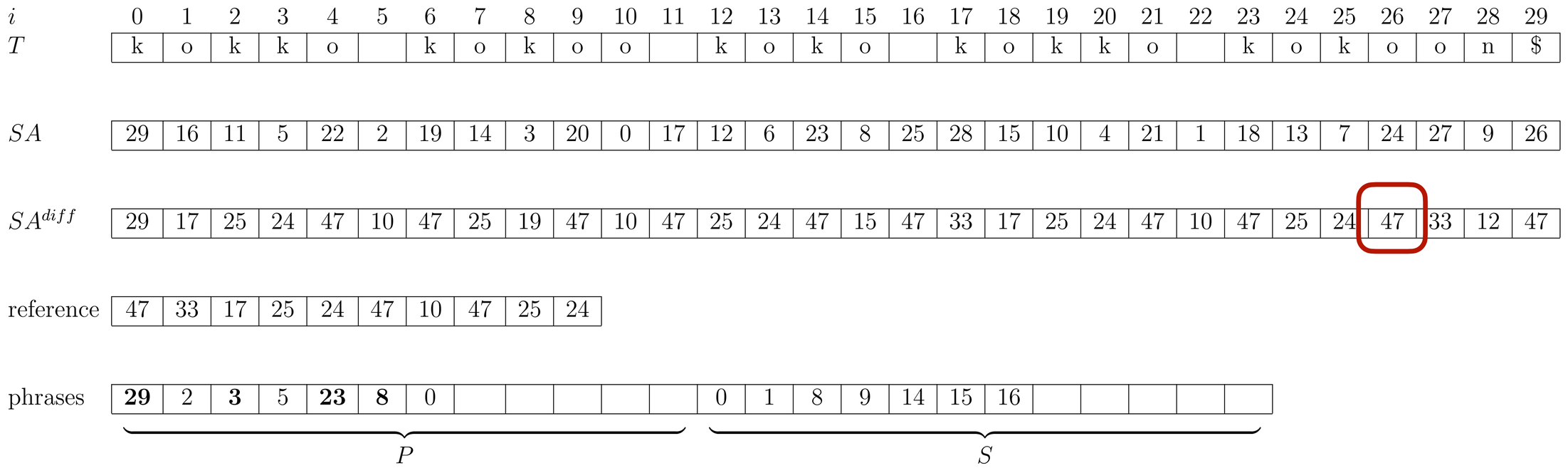
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 1 8 9 14 15
P
S

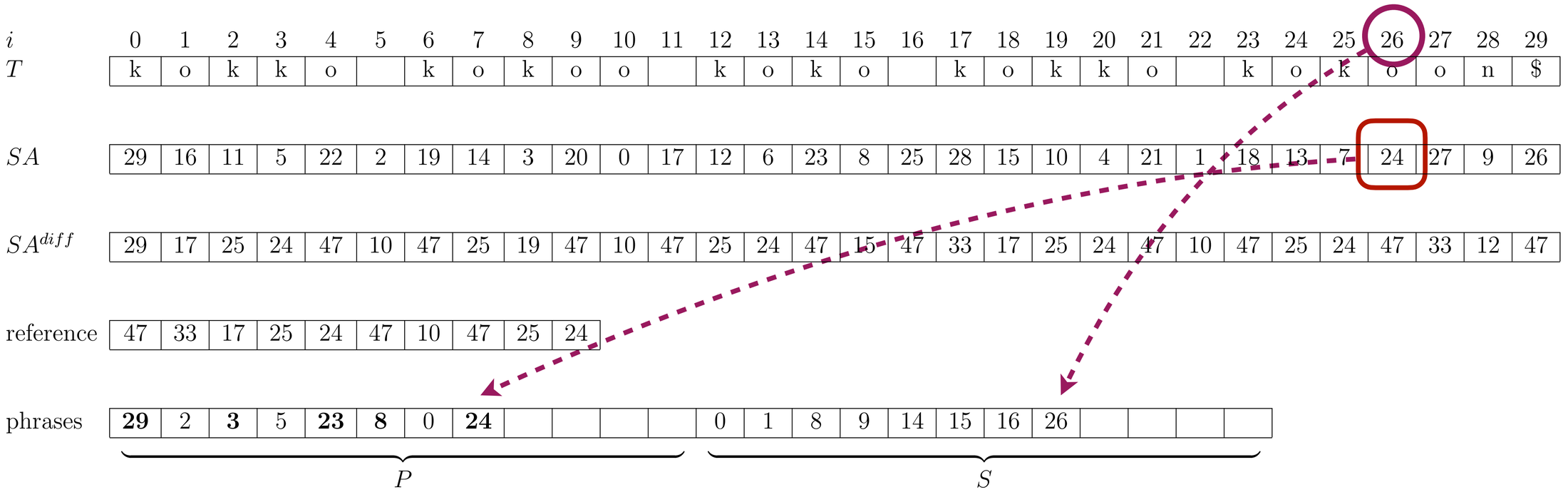
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k o k k o k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 0 1 8 9 14 15 16
P
S

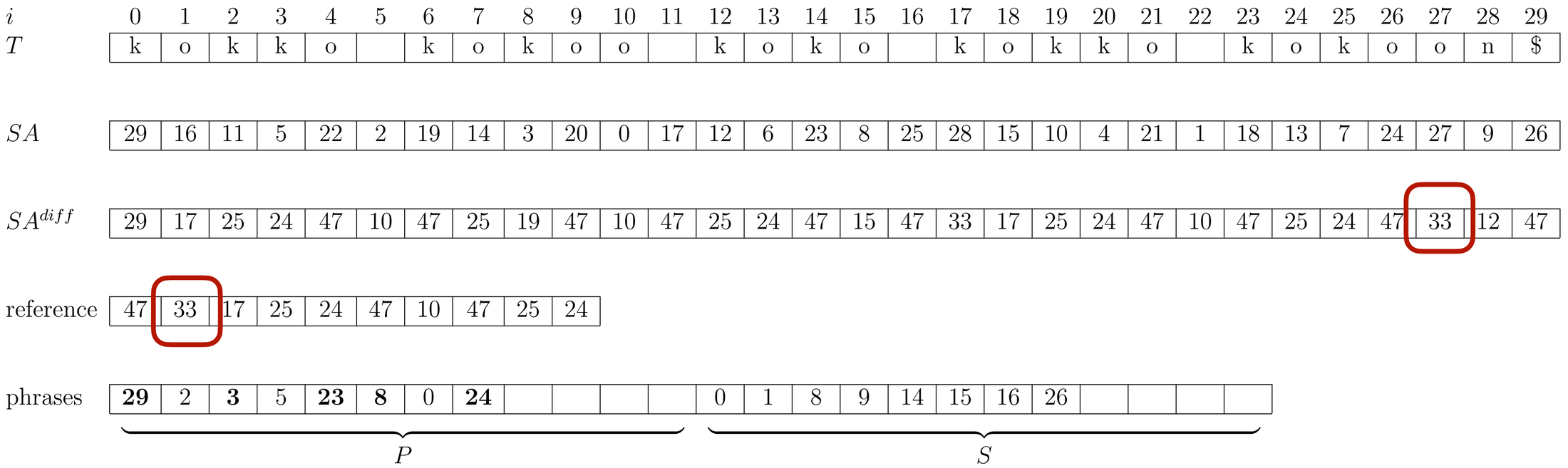
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k k o k k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 0 1 8 9 14 15 16
P
S

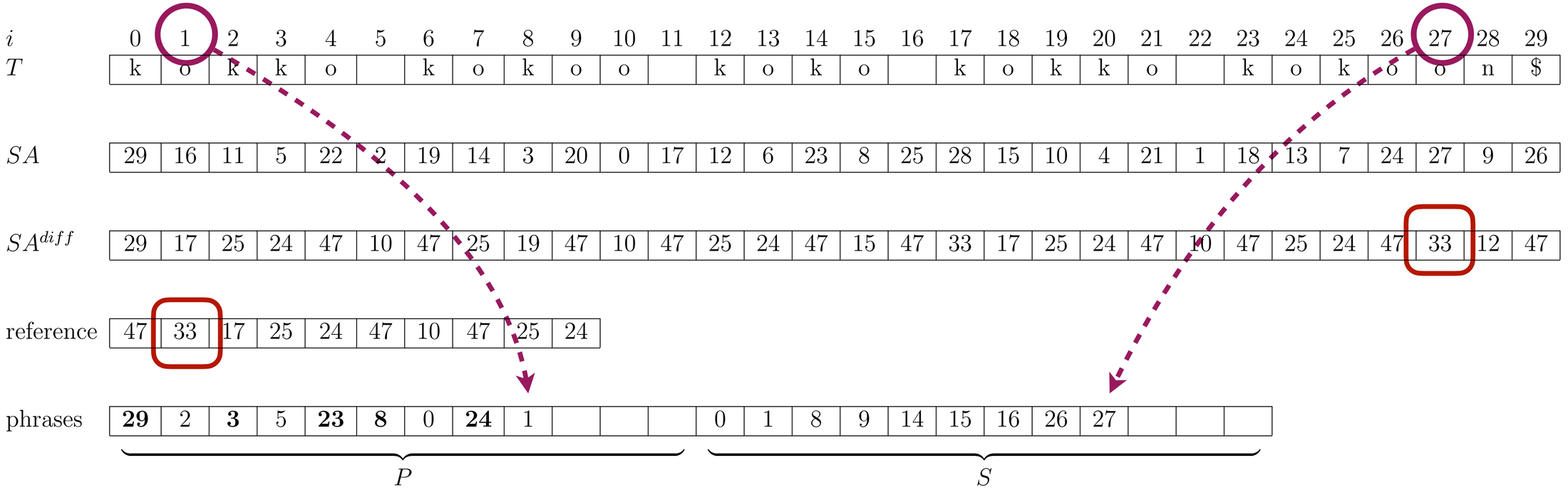
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o k k o k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 24 0 1 8 9 14 15 16 26
P
S

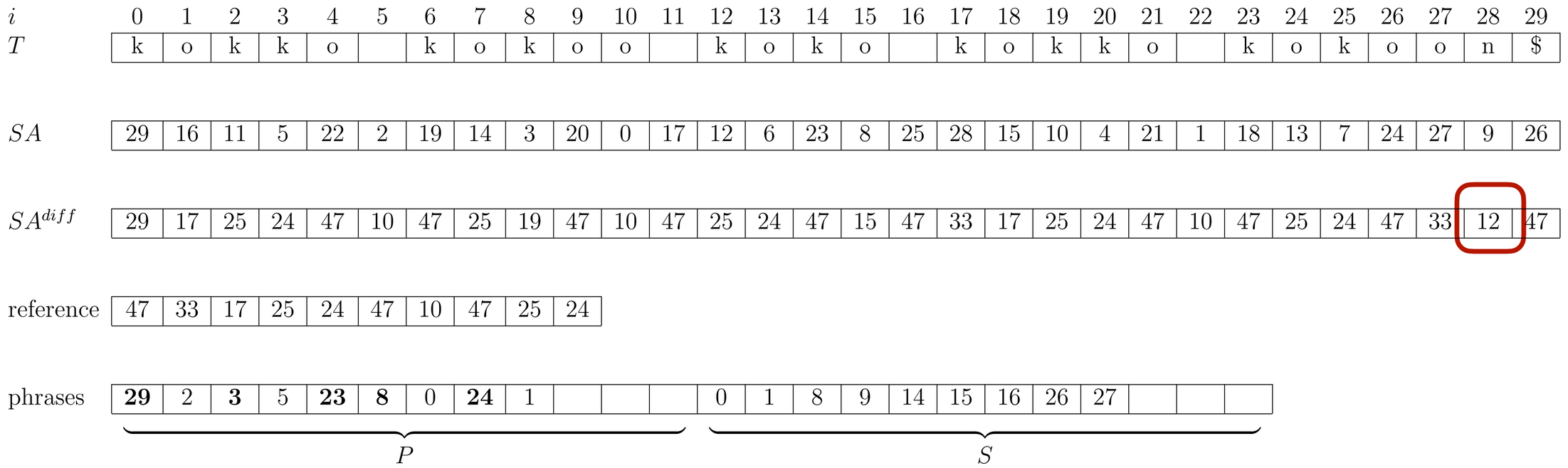
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 24 0 1 8 9 14 15 16 26
P
S

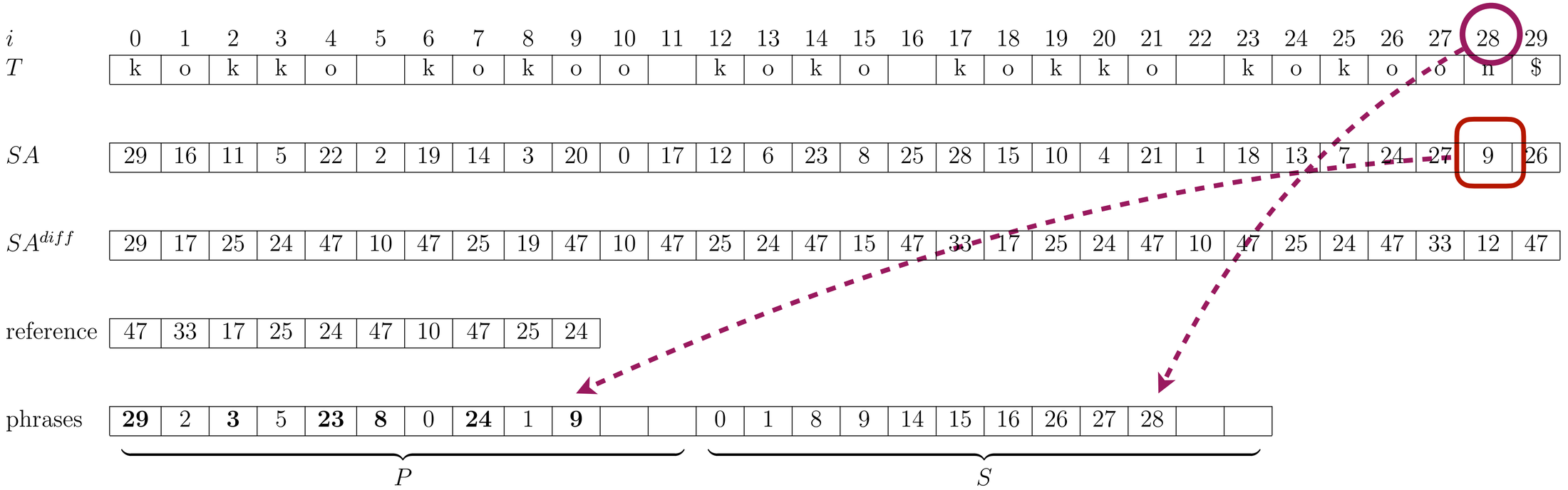
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 24 1 0 1 8 9 14 15 16 26 27
P
S

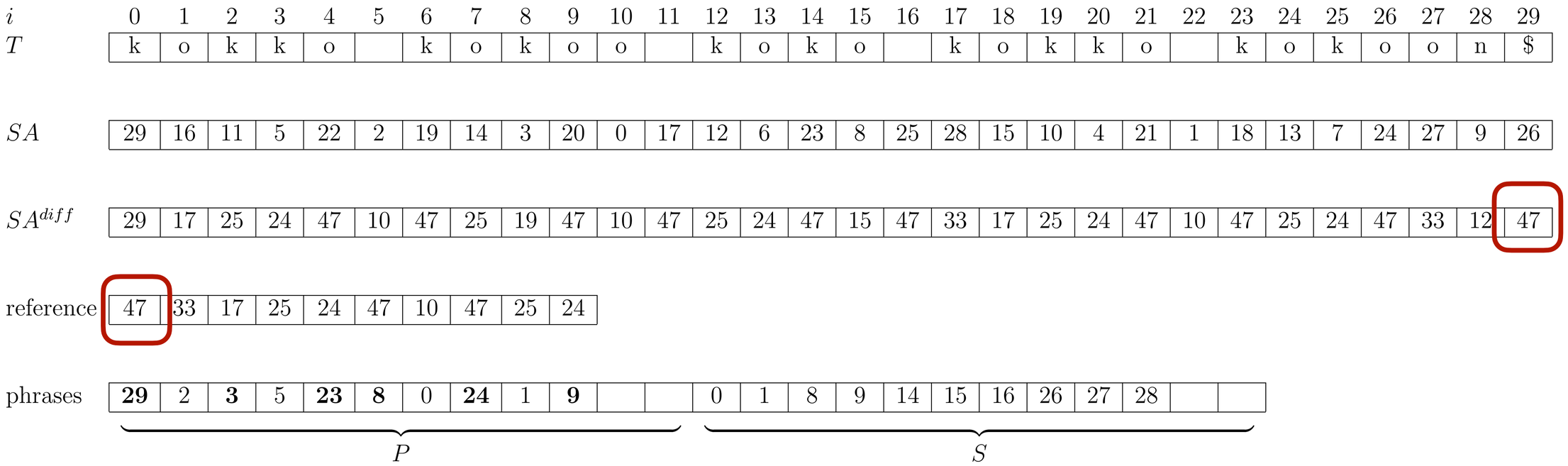
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 24 1 0 1 8 9 14 15 16 26 27
P
S

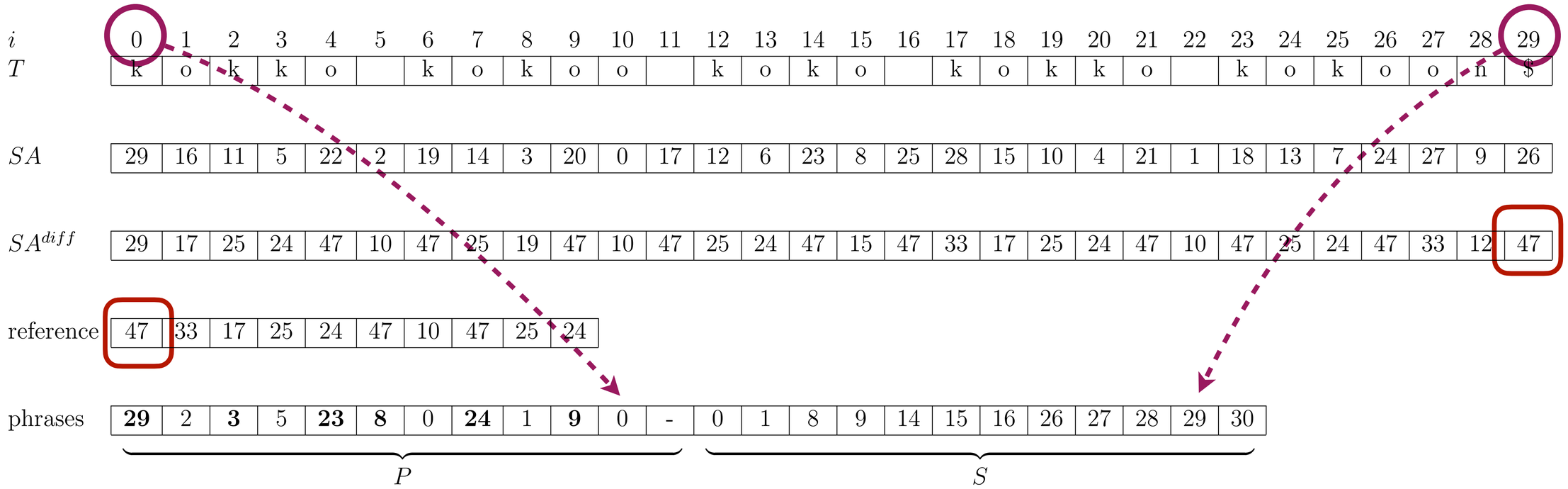
Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 24 1 9 0 1 8 9 14 15 16 26 27 28
P
S

Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k o k k o k o k k o k o o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 24 1 9 0 1 8 9 14 15 16 26 27 28
P
S

Example


i 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
T k o k k o k o k o o k o k o k k o k o n $
SA 29 16 11 5 22 2 19 14 3 20 0 17 12 6 23 8 25 28 15 10 4 21 1 18 13 7 24 27 9 26
SA^diff 29 17 25 24 47 10 47 25 19 47 10 47 25 24 47 15 47 33 17 25 24 47 10 47 25 24 47 33 12 47
reference 47 33 17 25 24 47 10 47 25 24
phrases 29 2 3 5 23 8 0 24 1 9 0 - 0 1 8 9 14 15 16 26 27 28 29 30
P
S

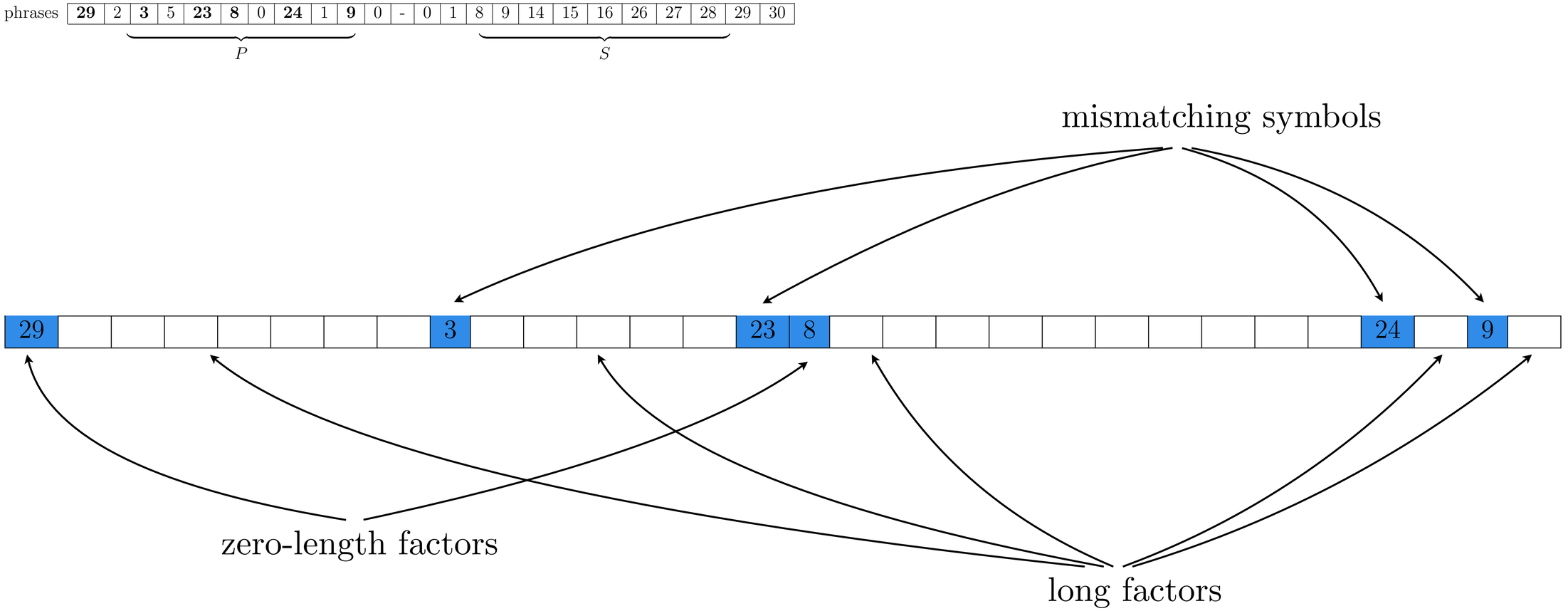
Example

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
T	k	o	k	k	o		k	o	k	o	o		k	o	k	o		k	o	k	k	o		k	o	k	o	o	n	$
SA	29	16	11	5	22	2	19	14	3	20	0	17	12	6	23	8	25	28	15	10	4	21	1	18	13	7	24	27	9	26
SA^{diff}	29	17	25	24	47	10	47	25	19	47	10	47	25	24	47	15	47	33	17	25	24	47	10	47	25	24	47	33	12	47

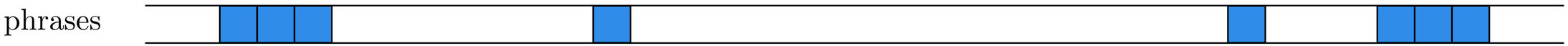
reference	47	33	17	25	24	47	10	47	25	24

phrases	**29**	2	**3**	5	**23**	**8**	0	**24**	1	**9**	0	-	0	1	8	9	14	15	16	26	27	28	29	30
	P												S											

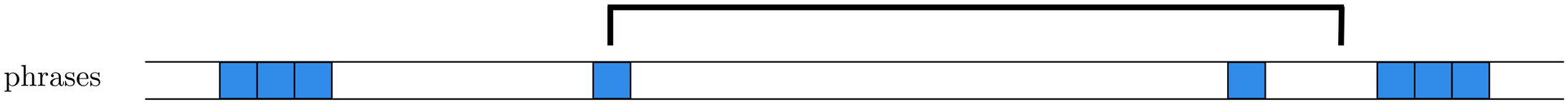
Example


phrases
29 2 3 5 23 8 0 24 1 9 0 - 0 1 8 9 14 15 16 26 27 28 29 30
P
S
mismatching symbols
29 3 23 8 24 9
zero-length factors
long factors

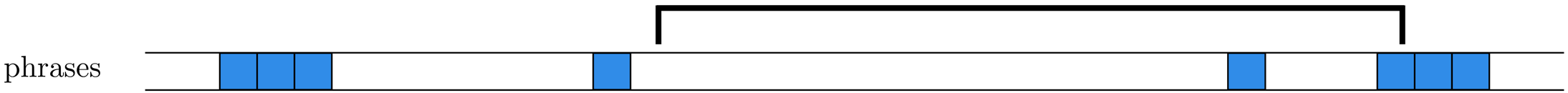
Example



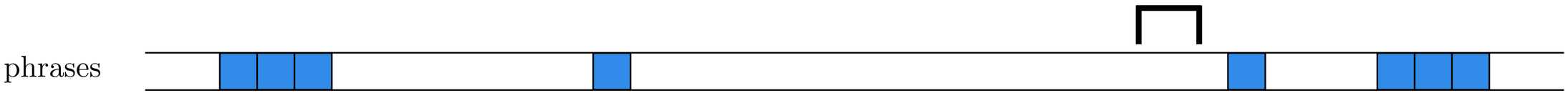
Example



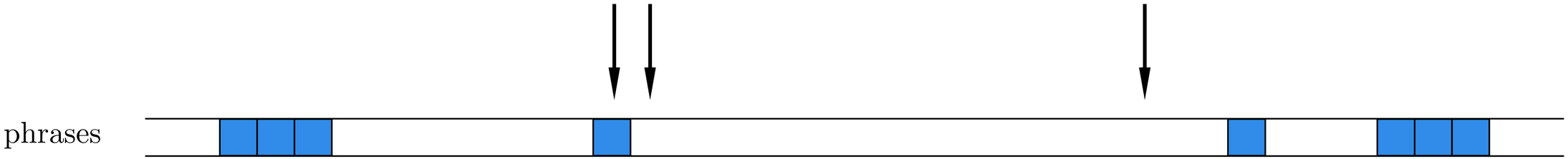
Example



Example

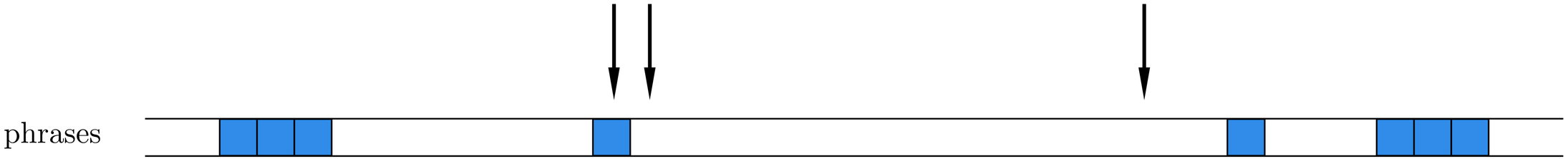


Example



- predecessor structure to find the phrase that contains the start of the SA interval

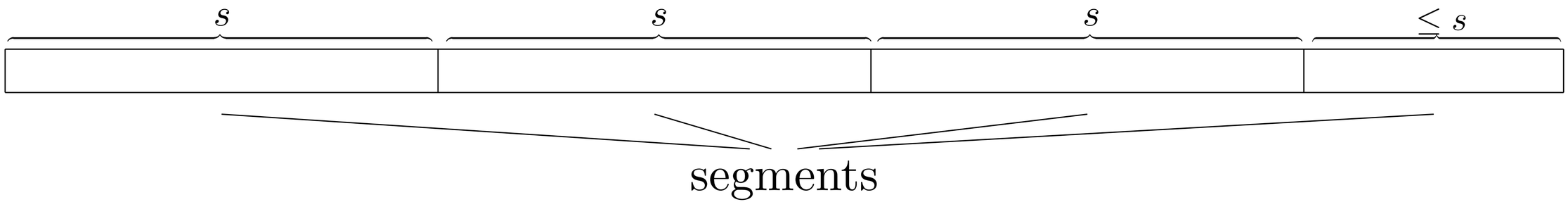
Example



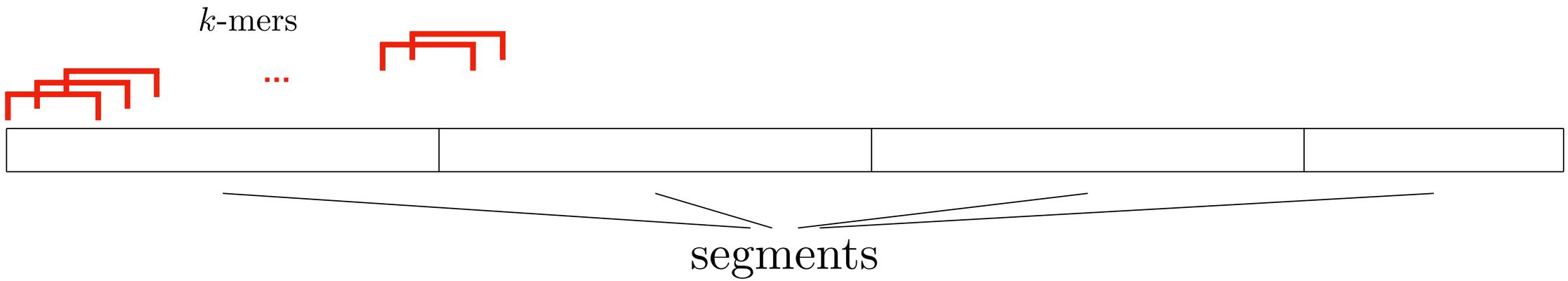
- predecessor structure to find the phrase that contains the start of the SA interval

- $value[i] = \begin{cases} P[i], & \text{if } phraseLength[i] = 1 \\ reference[\,P[i]\,] + prevSaValue - n, & \text{otherwise} \end{cases}$

Reference Generation

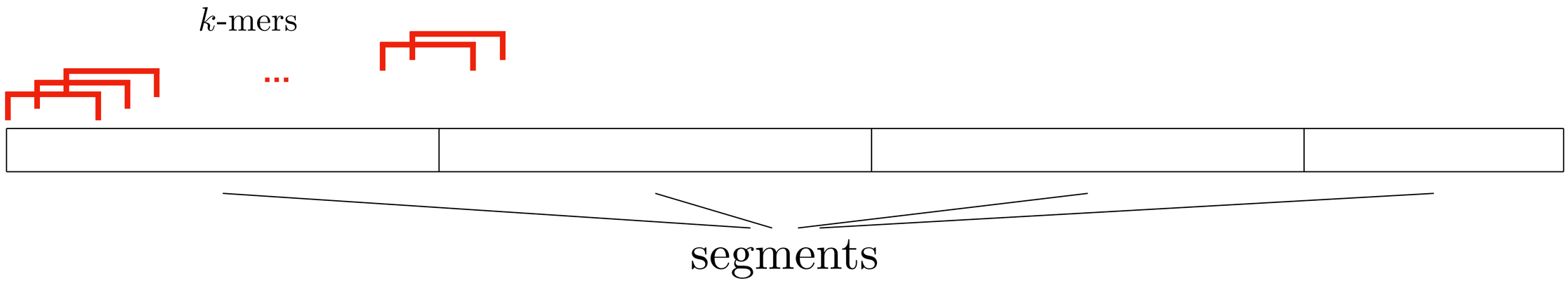


Reference Generation



$$\text{score}(seg_i) = (\Sigma_{x \in seg_i} f(x)^p)^{1/p}$$

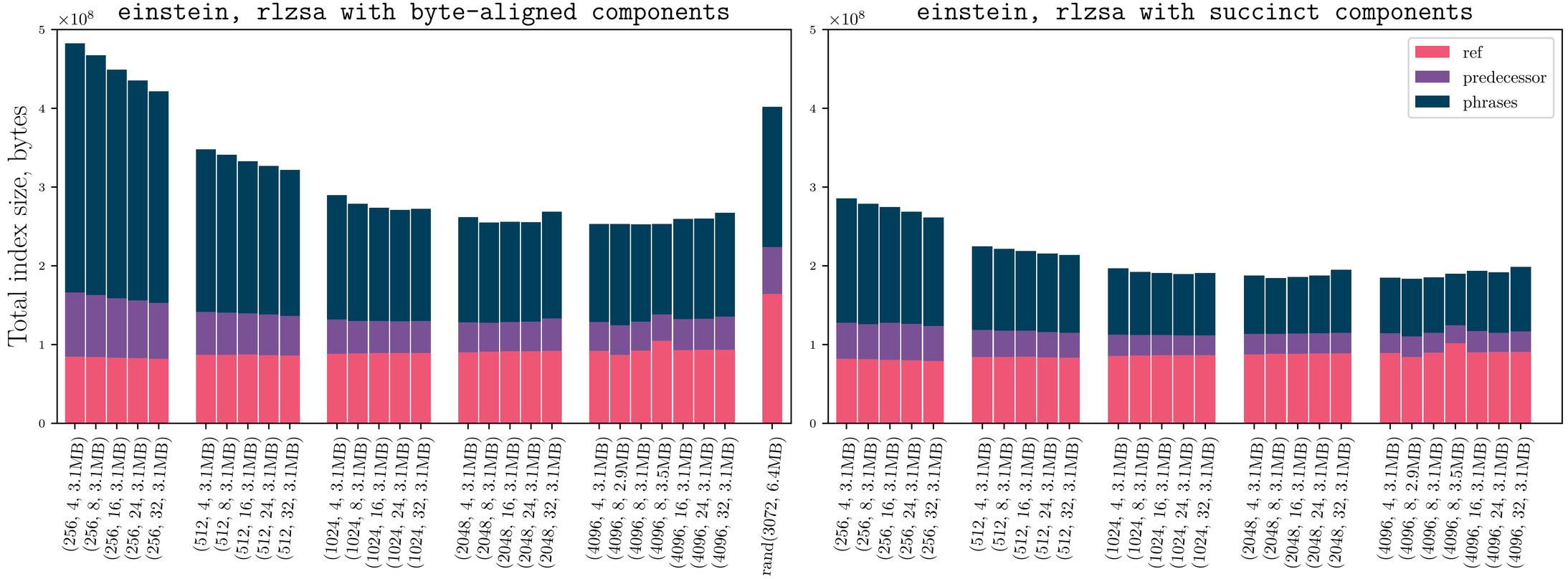
Reference Generation



$$\text{score}(seg_i) = (\Sigma_{x \in seg_i} f(x)^p)^{1/p}$$

- take the segment with the highest score, save its starting position
- for each k-mer from the chosen segment, reduce scores for all segments by corresponding k-mer frequency
- repeat until requested reference size is met
- sort starting positions, and write corresponding segments to the reference

Affect of (s, k) on Overall Index Size



Experiment

We compared our prototypes (*rlzsa* and *rlzsa-sdsl*) to other compressed indexes, replicating the experimental design used in the *r-index* paper (Gagie, Navarro, and Prezza, SODA 2018)

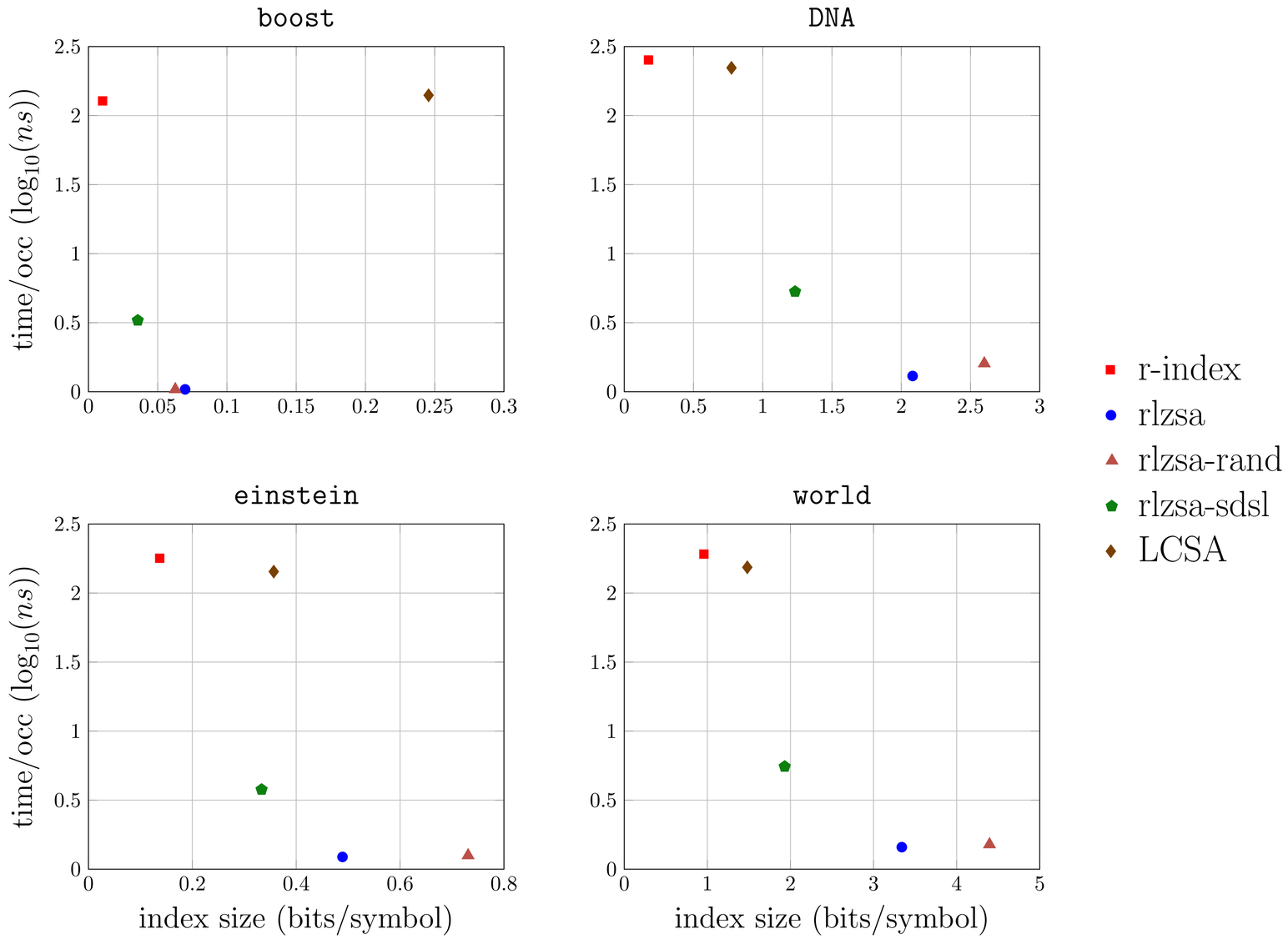
Datasets:

- boost — concatenated versions of GitHub's boost library — 600Mbyte
- DNA — concatenated copies of a DNA sequence of length 1000 with mutations — 600 Mbyte
- einstein — concatenated versions of Wikipedia's Einstein page — 600 Mbyte
- world — pdf files of CIA World Leaders from Jan 2003 to Dec 2009 — 45Mbyte

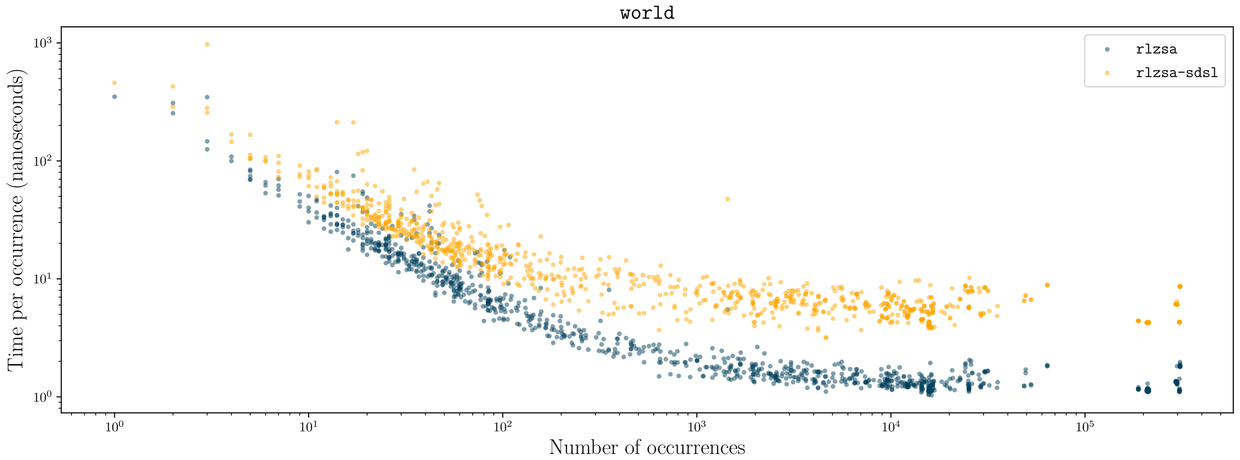
Search queries:

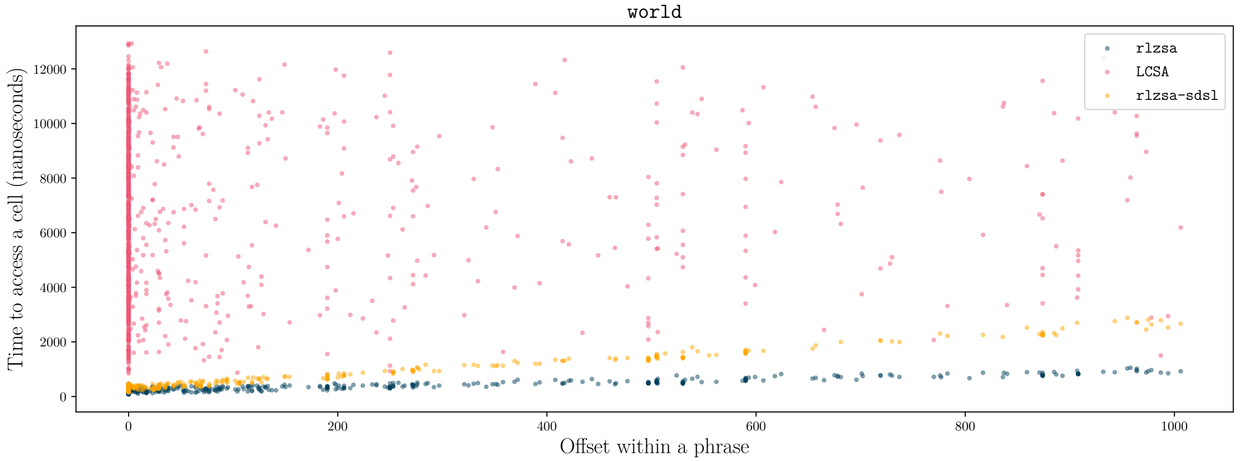
- 1000 patterns
- length = 8

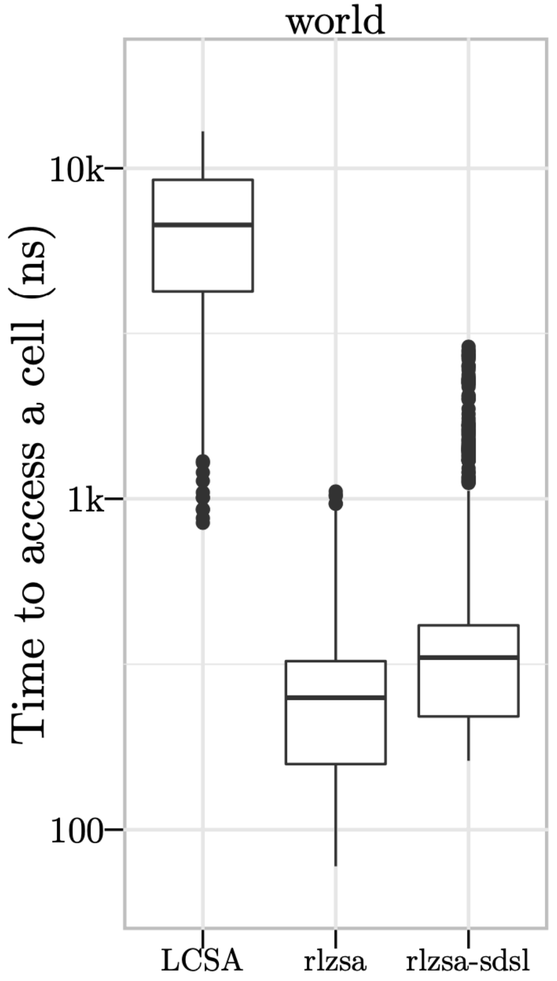
Experimental Results



Affect of Interval Size on Extraction Time







Future Work

- Apply it to
  - document (D) array (currently in submission)
  - ISA, LCP

- Best of both worlds?
  - Is there a way to derive a hybrid of the **r-index** and **rlzsa**? (we think, yes)

- Automatic choice of parameters (k, s, *ref*)

Thank you!