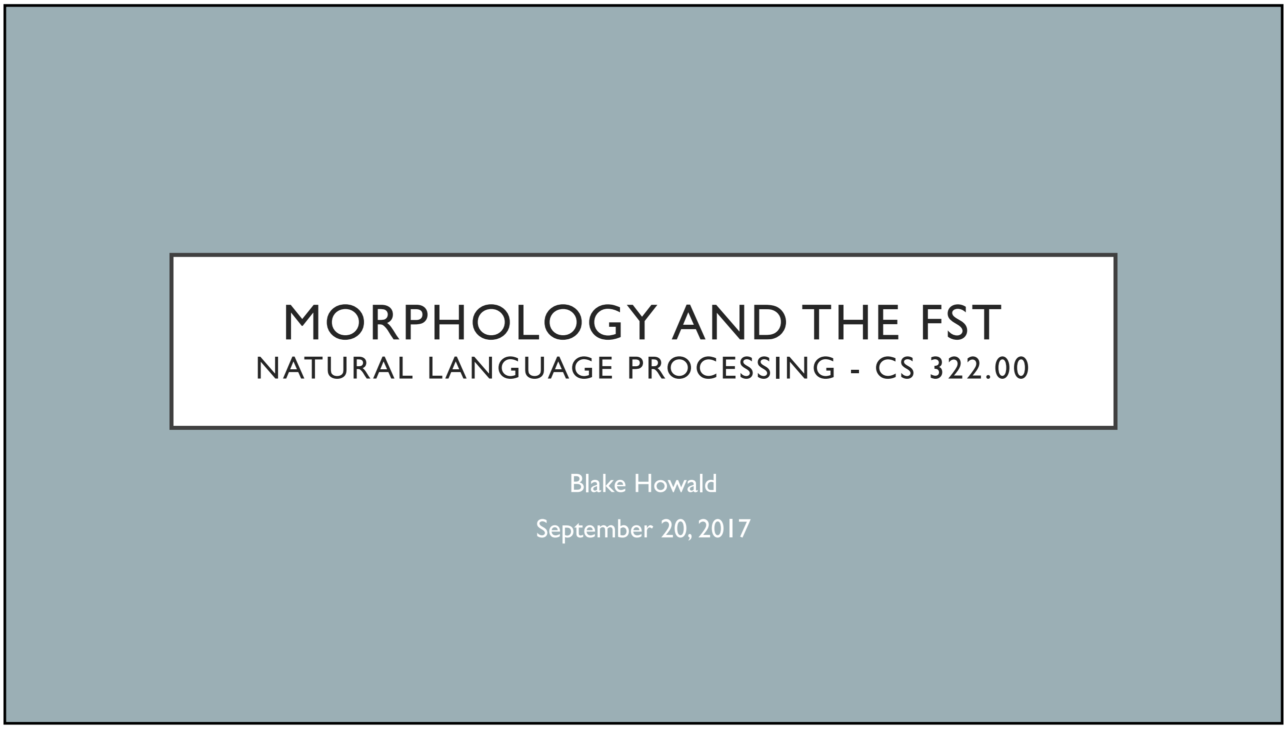
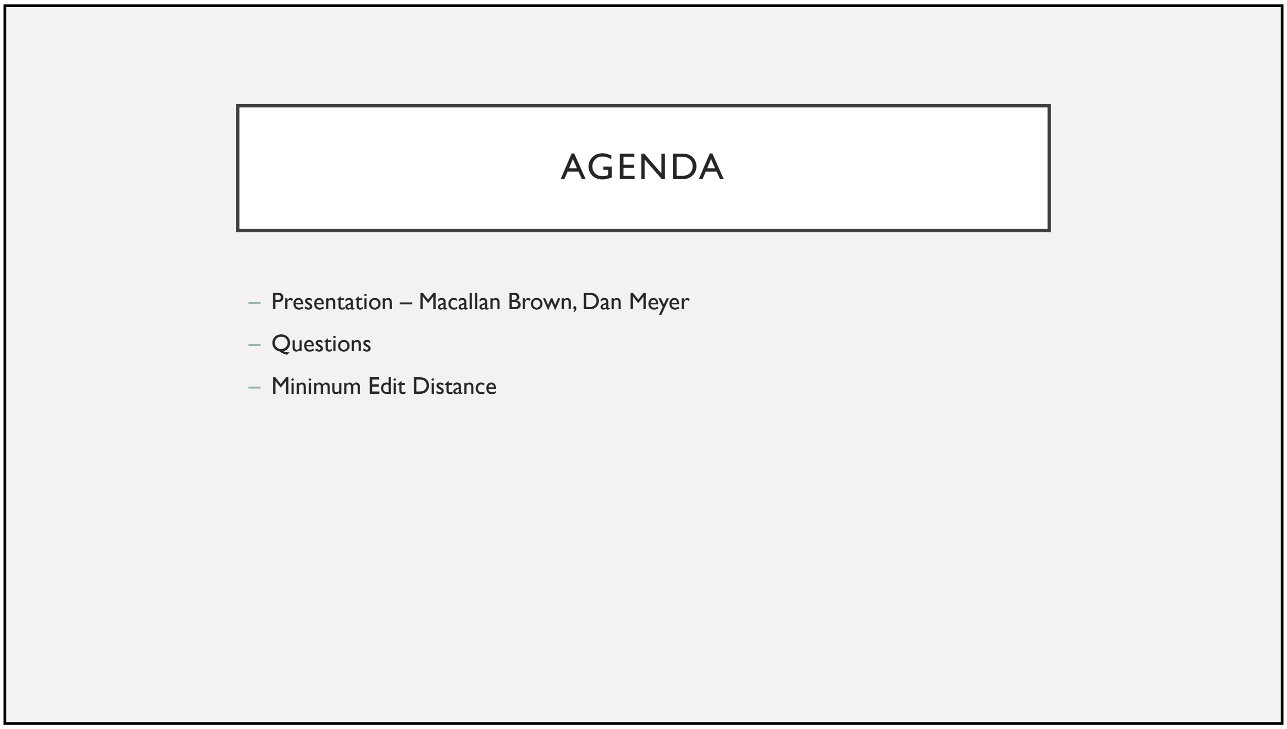


MORPHOLOGY AND THE FST

NATURAL LANGUAGE PROCESSING - CS 322.00

Blake Howald

September 20, 2017



AGENDA

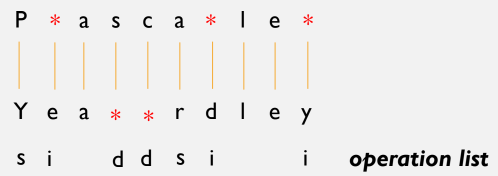
- Presentation – Macallan Brown, Dan Meyer
- Questions
- Minimum Edit Distance

MINIMUM EDIT DISTANCE

- Measure of string similarity – "how alike" are two strings
- Defined by the minimum number of editing operations needed to transform one string into another.
- Three operations available – *substitution, deletion, insertion*
- Cost associated with each operation – can be anything
  - Levenshtein Edit distance is typically 1 for deletion, insertion and 2 for substitution, but can be anything

MINIMUM EDIT DISTANCE

- Minimum edit distance dependent on *alignment*
  - *Pascale*
  - *Yeardley*



```
P * a s c a * l e *
| | | | | | | | | |
Y e a * * r d l e y
s i   d d s i     i
```

operation list

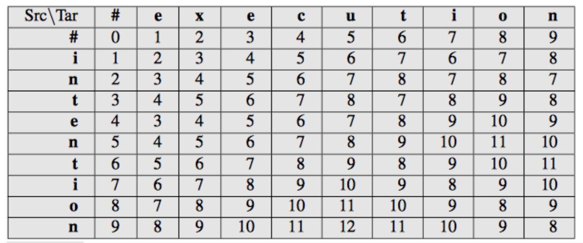
– Minimum Edit Distance = ?

MINIMUM EDIT DISTANCE

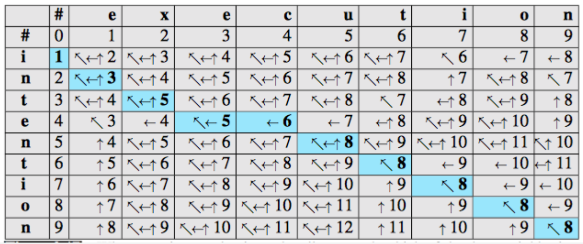
- Multiple alignments possible
- Multiple minimum alignments possible (alignments yielding the minimum edit distance)
    - Alignments will be important for speech recognition and machine translation
- Determining minimum edit distance (and all associated paths) requires construction minimum edit distance matrix

$$D[i,j] = \min \begin{cases} D[i-1,j]+1 \\ D[i,j-1]+1 \\ D[i-1,j-1]+\begin{cases} 2; & \text{if } source[i] \neq target[j] \\ 0; & \text{if } source[i] = target[j] \end{cases} \end{cases}$$

MINIMUM EDIT DISTANCE



Src\Tar	#	e	x	e	c	u	t	i	o	n
#	0	1	2	3	4	5	6	7	8	9
i	1	2	3	4	5	6	7	6	7	8
n	2	3	4	5	6	7	8	7	8	7
t	3	4	5	6	7	8	7	8	9	8
e	4	3	4	5	6	7	8	9	10	9
n	5	4	5	6	7	8	9	10	11	10
t	6	5	6	7	8	9	8	9	10	11
i	7	6	7	8	9	10	9	8	9	10
o	8	7	8	9	10	11	10	9	8	9
n	9	8	9	10	11	12	11	10	9	8



	#	e	x	e	c	u	t	i	o	n
#	0	1	2	3	4	5	6	7	8	9
i	**1**	↖←↑ 2	↖←↑ 3	↖←↑ 4	↖←↑ 5	↖←↑ 6	↖←↑ 7	↖ 6	← 7	← 8
n	2	↖←↑ **3**	↖←↑ 4	↖←↑ 5	↖←↑ 6	↖←↑ 7	↖←↑ 8	↑ 7	↖←↑ 8	↖ 7
t	3	↖←↑ 4	↖←↑ **5**	↖←↑ 6	↖←↑ 7	↖←↑ 8	↖ 7	←↑ 8	↖←↑ 9	↑ 8
e	4	↖ 3	← 4	↖← **5**	← **6**	← 7	←↑ 8	↖←↑ 9	↖←↑ 10	↑ 9
n	5	↑ 4	↖←↑ 5	↖←↑ 6	↖←↑ 7	↖←↑ **8**	↖←↑ 9	↖←↑ 10	↖←↑ 11	↖↑ 10
t	6	↑ 5	↖←↑ 6	↖←↑ 7	↖←↑ 8	↖←↑ 9	↖ **8**	← 9	← 10	←↑ 11
i	7	↑ 6	↖←↑ 7	↖←↑ 8	↖←↑ 9	↖←↑ 10	↑ 9	↖ **8**	← 9	← 10
o	8	↑ 7	↖←↑ 8	↖←↑ 9	↖←↑ 10	↖←↑ 11	↑ 10	↑ 9	↖ **8**	← 9
n	9	↑ 8	↖←↑ 9	↖←↑ 10	↖←↑ 11	↖←↑ 12	↑ 11	↑ 10	↑ 9	↖ **8**

- Dynamic Programming (table driven algorithm)
- Maintenance of 'back trace' to minimum edit path(s)
    - Contiguous cells indicate deletion (here horizontal), insertion (here, vertical), substitution (diagonal in all cases)
- Minimum edit distance is cell [source length, target length] ([n, m])