SFU NatLangLab

CMPT 413: Computational Linguistics

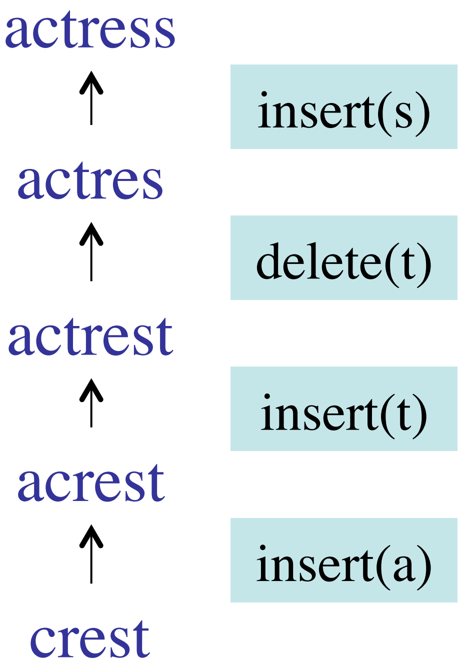
ED1: Introducing Edit Distance

Anoop Sarkar

http://www.cs.sfu.ca/~anoop

Minimum Cost Edit Distance

- Edit a source string into a target string
- Each edit has a cost
- Find the minimum cost edit(s)



actress

↑ insert(s)

actres

↑ delete(t)

actrest

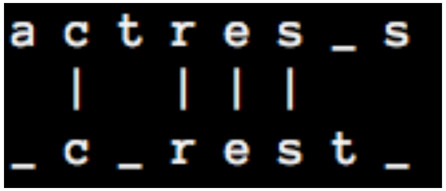
↑ insert(t)

acrest

↑ insert(a)

crest

Minimum Cost Edit Distance



```
a c t r e s _ s
  |   | | |
_ c _ r e s t _
```

target

source

actress

↑

actres

↑

actrest

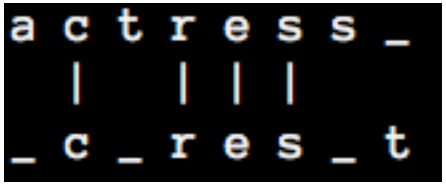
↑

acrest

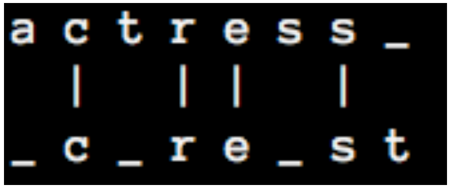
↑

crest

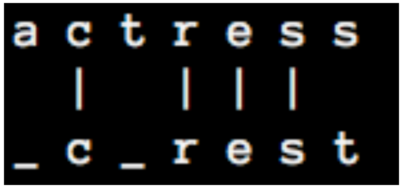
minimum cost edit distance can be accomplished in multiple ways



```
a c t r e s s _
  |   | | |
_ c _ r e s _ t
```



```
a c t r e s s _
  |   | |   |
_ c _ r e _ s t
```



```
a c t r e s s
  |   | | |
_ c _ r e s t
```

Only 4 ways to edit source to target **for this pair**

Levenshtein Distance

- Cost is fixed across characters
  - Insertion cost is 1
  - Deletion cost is 1
- Two different costs for substitutions
  - Substitution cost is 1 (transformation)
  - Substitution cost is 2 (one deletion + one insertion)

Левенштейн Владимир

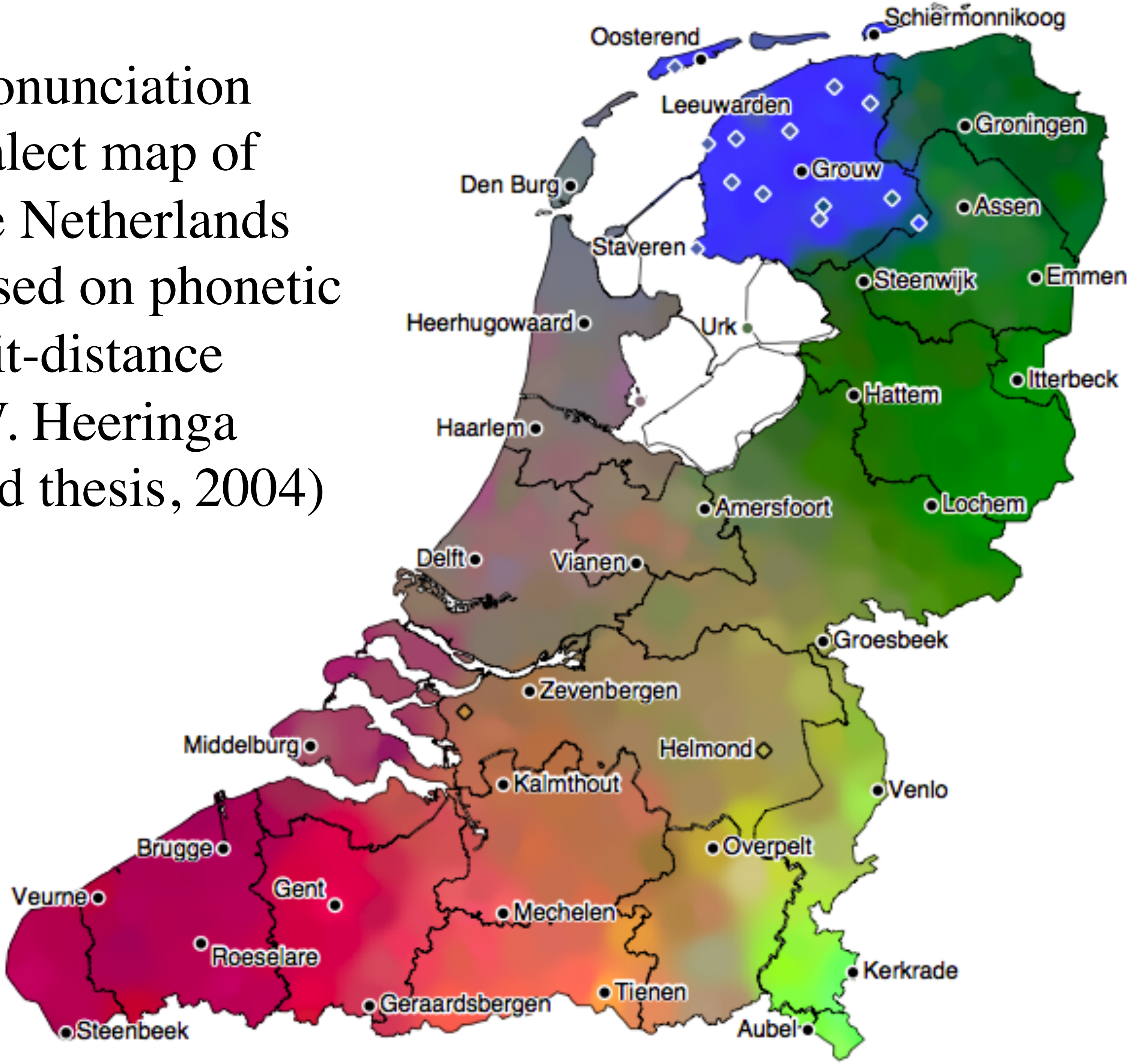
Vladimir Levenshtein

What's the edit distance?

Edit distance

- Useful in many NLP applications
- In some cases, we need edits with multiple characters, e.g. 2 chars deleted for one cost
- Comparing system output with human output, e.g. *input:* ibm *output:* IBM vs. Ibm (TrueCasing of speech recognition output)
- Error correction, e.g. spelling correction
- Defined over character edits or word edits, e.g. MT evaluation:
  - Foreign investment in Jiangsu ‘s agriculture on the increase
  - Foreign investment in Jiangsu agricultural investment increased

Pronunciation dialect map of the Netherlands based on phonetic edit-distance (W. Heeringa Phd thesis, 2004)


Schiermonnikoog
Oosterend
Leeuwarden
Groningen
Grouw
Den Burg
Assen
Staveren
Steenwijk
Emmen
Heerhugowaard
Urk
Itterbeck
Hattem
Haarlem
Amersfoort
Lochem
Delft
Vianen
Groesbeek
Zevenbergen
Middelburg
Helmond
Kalmthout
Venlo
Brugge
Overpelt
Gent
Veurne
Mechelen
Roeselare
Kerkrade
Tienen
Geraardsbergen
Aubel
Steenbeek