Measuring linguistic distances
Phonetic distance - Levenshtein distance

Sebastian Kürschner, University of Groningen

Levels of measuring linguistic distance

- linguistic distances can be measured on different linguistic levels
  - lexicon: how many words are cognate?
  - phonetics/phonology: how much phonetic distance is there between cognates?
  - morphology: what is expressed where and how, and how similar are the languages in this respect?
  - syntax: to what extent are the syntactic systems similar?

Phonetic distance: Levenshtein-afstand

- computational method for comparison of related language varieties
- mostly used for measuring phonetic differences (Heeringa 2004)
- string mapping: comparing two strings
  - the costs of the least operations necessary for mapping are calculated
  - operations are insertions, deletions, and substitutions
  - can be normalized by word length

Levenshtein distance: example

Danish 

hjemme

– Swedish 

hemma

'at home'

Danish 

guld

– Swedish 

guld

'gold'

Levenshtein distance: example

<table>
<thead>
<tr>
<th></th>
<th>j</th>
<th>c</th>
<th>m</th>
<th>ø</th>
<th>h</th>
<th>e</th>
<th>m:</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Danish 

guld

– Swedish 

guld

'gold'

<table>
<thead>
<tr>
<th></th>
<th>g</th>
<th>u</th>
<th>l</th>
<th>g</th>
<th>o</th>
<th>l</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculation of Levenshtein distance

1. Matching the strings
   - Find matching sounds and align them
   - align the non-matching sounds
   - often the system is informed about the vowel-/consonant-distinction to make likely matchings according to syllable structure
   - i.e. find matching vowels and consonants and align them
   - align non-matching vowels with vowels and non-matching consonants with consonants only
2. Calculating the distance

Calculation of Levenshtein distance

- based on phonetic transcriptions:
  - simplest method: each difference is counted i.e. also [a] vs. [a:]
  - if necessary, difference between sounds can be weighed according to similarity e.g. only 0.5 for [m] vs. [m:]
- based on feature systems
  - difference is calculated according to difference in phonetic features
    [a] and [e] are different to a smaller degree than [a] and [i]
Calculation of distance between varieties

› $k$ word pairs, consisting each of two representations of the same word in two varieties

› calculate Levenshtein distance for each of the $k$ pairs

› distance between varieties = average of the $k$ distances

Hypothesis

The phonetic distance of two languages cannot exceed a certain degree for mutual intelligibility to be possible