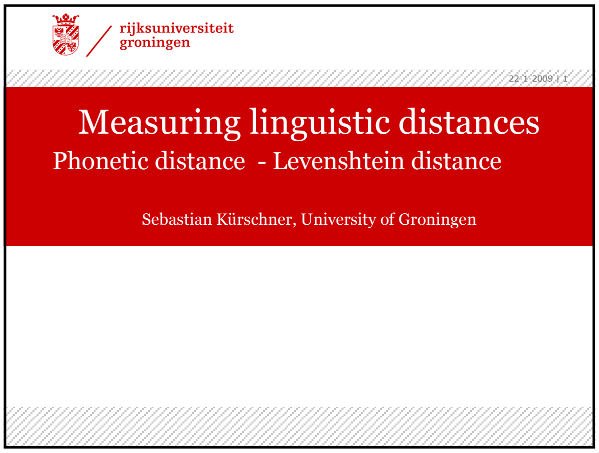
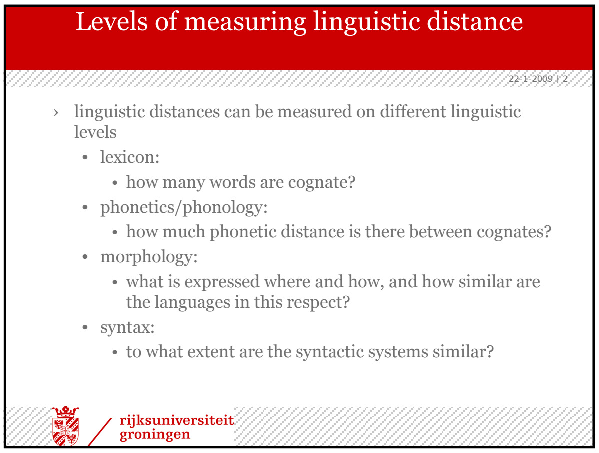


Measuring linguistic distances

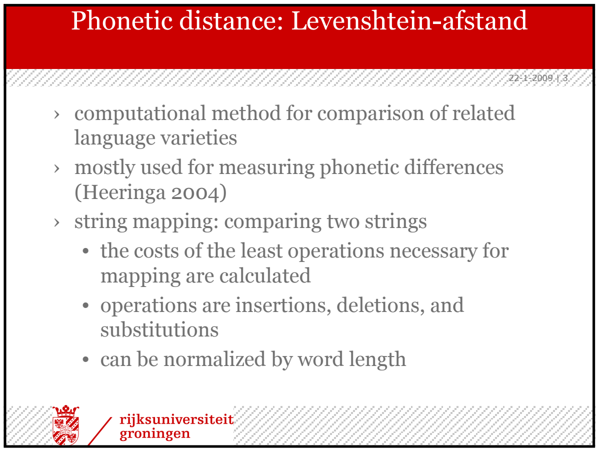
Phonetic distance - Levenshtein distance

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?

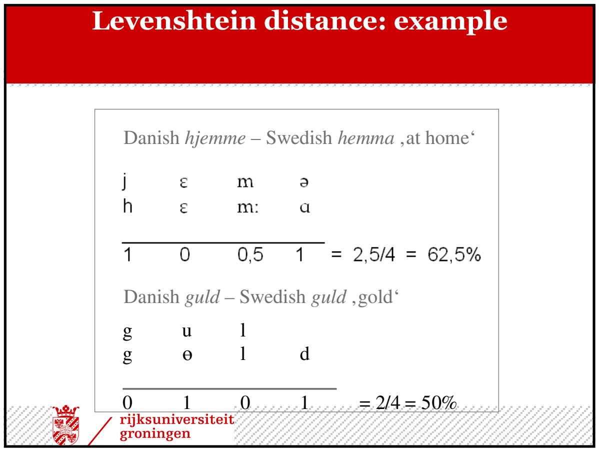


Phonetic distance: Levenshtein-afstand

› 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



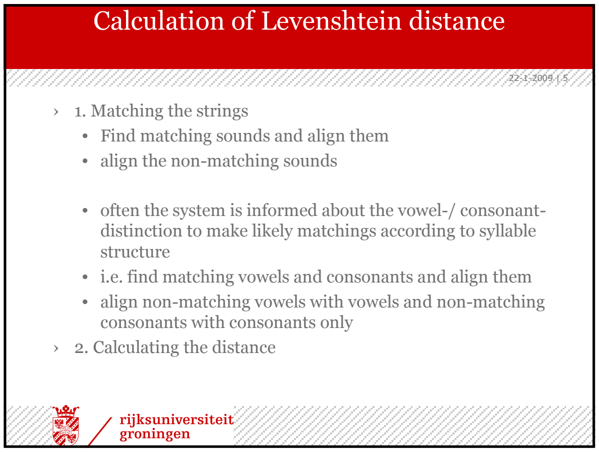
Levenshtein distance: example

Danish *hjemme* – Swedish *hemma* ‚at home‘

j	ɛ	m	ə	
h	ɛ	mː	ɑ	
1	0	0,5	1	= 2,5/4 = 62,5%

Danish *guld* – Swedish *guld* ‚gold‘

g	u	l		
g	ɵ	l	d	
0	1	0	1	= 2/4 = 50%

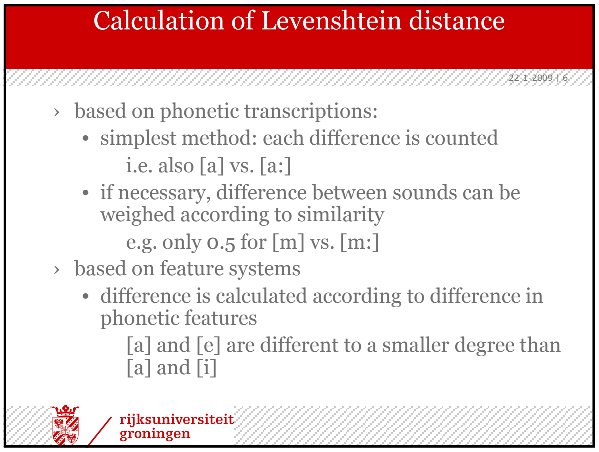


Calculation of Levenshtein distance

› 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

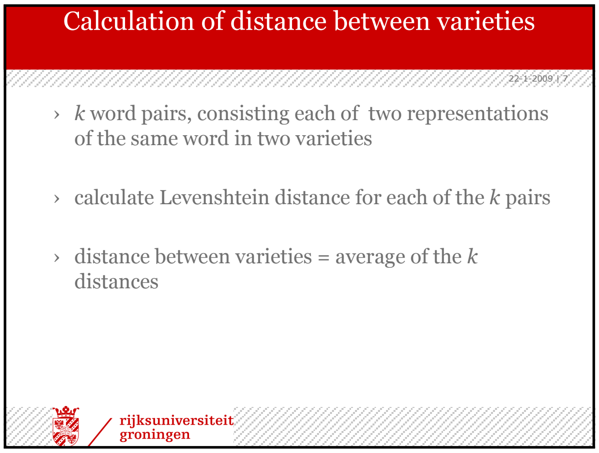
› based on phonetic transcriptions:
  - simplest method: each difference is counted

    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]

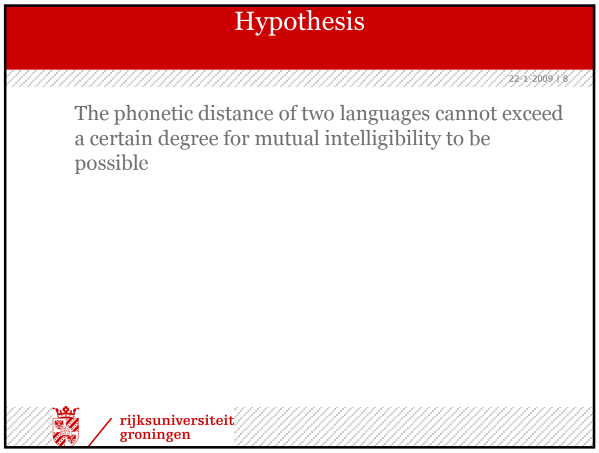


Calculation of distance between varieties

22-1-2009 | 7

- *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

rijksuniversiteit
groningen



Hypothesis

22-1-2009 | 8

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

rijksuniversiteit
groningen