# LED-A: a web app for measuring distances in the sound components among local dialects

Wilbert Heeringa & Vincent van Heuven & Hans Van de Velde

FRYSKE AKADEMY

Methods XVII

Mainz, August 1, 2022

# Introduction

- We present a new app for measuring dialect variation with Levenshtein distance.
- Proposed by Vladimir Levenshtein in 1965.
- Introduced in dialectology by Brett Kessler in 1995.
- He measured linguistic distances among Irish Gaelic dialects.
- Later widely used for many other language varieties.
- Calculate the cost of changing one string of characters into another.

# Levenshtein distance

- Example: *milk* is pronounced as [mɛlək] in the dialect of Haarlem and as [mɔlkə] in the dialect of Grouw.
- Find cheapest mapping of [mɛlək] → [mɔlkə]:

| cumulative maxim. cost | 1 | 2 | 3 | 3.5 | 4.5 | 5 |
|---|---|---|---|---|---|---|
| | m | ɛ | l | ə | k | |
| | m | ɔ | l | | k | ə |
| actual cost | | 1 | | 0.5 | | 0.5 |

- Raw distance: $1 + 0.5 + 0.5 = 2$ Normalized distance: $2 \, / \, 5 = 0.4 = 40\%$

# Aggregated distance

- Example: calculate pronunciation differences between Grouw and Haarlem dialects for 6 words:

| | Grouw | Haarlem | cost | max. | norm. |
|---|---|---|---|---|---|
| | | | | cost | cost |
| work | ʋʏrk | ʋɛrək | 1.5 | 4.5 | 0.33 |
| ship | skɪp | sxɪp | 1 | 4 | 0.25 |
| finger | fɪŋər | vɪŋər | 1 | 5 | 0.2 |
| wine | ʋin | ʋɛin | 0.5 | 3.5 | 0.14 |
| house | huz | hœys | 0.5 | 3.5 | 0.14 |
| milk | mɔlkə | mɛlək | 2 | 5 | 0.4 |
| | | | 6.5 | | 1.46 |

- Raw distance:
  $6.5/6 = 1.08$

  Normalized distance:
  $1.46/6 = 0.243 = 24.3\%$

# Software

- Visual Dialectometry (VDM).
- DiaTech: related to VDM, but it is a webapp and includes also Levenshtein distance.
- R*u*G/L$^{04}$: set of functions to be entered as command line commands using a keyboard.
- Gabmap: webapp at `gabmap.nl`, Docker version at `https://github.com/pebbe/Gabmap-docker`.
- Python packages: editdistance, LingPy.
- R packages: iL04 (`http://www.let.rug.nl/~kleiweg/L04/R/`), stringdist, alineR, dialectR (at GitHub).

- Goal: make dialectometry as easy as possible:
  - **L**evenshtein **E**dit **D**istance **A**pp,
  - refers to 'Lampje' or Gyro Gearloose's ‘Little Helper’ in Donald Duck comics; ‘Lampje’ is Dutch for Little Lamp (LED = Light Emitting Diode = lamp).
- Availability:
  - `https://www.led-a.org`
  - example data sets under 'Examples'.

Pictures of Gyro and Helper are taken from `https://donaldduck.fandom.com/nl/wiki/Willie_Wortel` and `https://donaldduck.fandom.com/nl/wiki/Lampje`, CC BY-SA 3.0 license.

# Distance measures

- Binary item comparison (Séguy 1973)
- Levenshtein distance plain

  indel is 0.5 or 1
- Levenshtein distance IPA feature-based (Almeida & Braun)

  indel scaled between 0 and 0.5 or joint scaling of subsitutions and indels between 0 and 1.
- Levenshtein distance PMI-based

  Wieling, Prokić & Nerbonne (2009), Wieling (2012)

# VC-sensitive

The minimum cost is based on an alignment in which:

- a vowel matches with a vowel
- a consonant matches with a consonant

Optionally the user can allow:

- [j̊], [w], [i] or [u] to match with anything
- [ə] (schwa) or [ɐ] to match with a sonorant

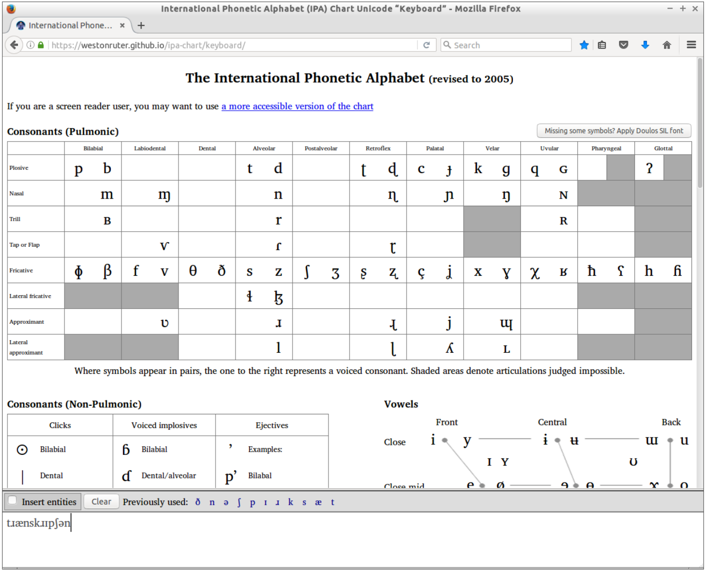
Bolognesi & Heeringa (2002), Heeringa (2004), Wieling et al. (2009)

# Levenshtein

- Both ‘raw’ and normalized Levenshtein distances can be calculated.
- It is possible to measure distances due to differences in only vowels or only consonants (indels or substitutions).

# Input: transcriptions

- Excel sheet (Microsoft Excel/LibreOffice Calc).
- The transcriptions should be in IPA Unicode (use `https://westonruter.github.io/ipa-chart/keyboard/`)
- Multiple transcriptions per item are possible.


International Phonetic Alphabet (IPA) Chart Unicode “Keyboard” - Mozilla Firefox
International Phone…
https://westonruter.github.io/ipa-chart/keyboard/
Search

## The International Phonetic Alphabet (revised to 2005)

If you are a screen reader user, you may want to use a more accessible version of the chart

### Consonants (Pulmonic)

Missing some symbols? Apply Doulos SIL font

| | Bilabial | Labiodental | Dental | Alveolar | Postalveolar | Retroflex | Palatal | Velar | Uvular | Pharyngeal | Glottal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plosive | p b | | | t d | | ʈ ɖ | c ɟ | k g | q ɢ | ▨ | ʔ ▨ |
| Nasal | m | ɱ | | n | | ɳ | ɲ | ŋ | ɴ | ▨ | ▨ |
| Trill | ʙ | | | r | | | | ▨ | ʀ | | ▨ |
| Tap or Flap | | ⱱ | | ɾ | | ɽ | | ▨ | | | ▨ |
| Fricative | ɸ β | f v | θ ð | s z | ʃ ʒ | ʂ ʐ | ç ʝ | x ɣ | χ ʁ | ħ ʕ | h ɦ |
| Lateral fricative | ▨ | ▨ | | ɬ ɮ | | | | | | ▨ | ▨ |
| Approximant | | ʋ | | ɹ | | ɻ | j | ɰ | | | ▨ |
| Lateral approximant | ▨ | ▨ | | l | | ɭ | ʎ | ʟ | | ▨ | ▨ |

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

### Consonants (Non-Pulmonic)

| Clicks | | Voiced implosives | | Ejectives | |
|---|---|---|---|---|---|
| ʘ | Bilabial | ɓ | Bilabial | ʼ | Examples: |
| ǀ | Dental | ɗ | Dental/alveolar | pʼ | Bilabal |

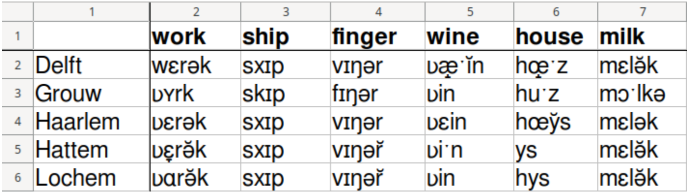
### Vowels

Previously used: ð n ə ʃ p ɪ ɹ k s æ t

tɹænskɹɪpʃən

# Format (1)

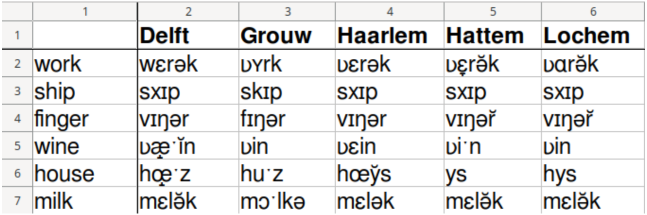
Rows are locations, columns are items



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | | **work** | **ship** | **finger** | **wine** | **house** | **milk** |
| 2 | Delft | wɛrək | sxɪp | vɪŋər | ʋæ̞ˑĭn | hœ̞ˑz | mɛlə̆k |
| 3 | Grouw | ʋʏrk | skɪp | fɪŋər | ʋin | huˑz | mɔˑlkə |
| 4 | Haarlem | ʋɛrək | sxɪp | vɪŋər | ʋɛin | hœy̆s | mɛlək |
| 5 | Hattem | ʋɛ̞rə̆k | sxɪp | vɪŋəř | ʋiˑn | ys | mɛlə̆k |
| 6 | Lochem | ʋɑrə̆k | sxɪp | vɪŋəř | ʋin | hys | mɛlə̆k |

# Format (2)

Rows are items, columns are locations



|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 |  | **Delft** | **Grouw** | **Haarlem** | **Hattem** | **Lochem** |
| 2 | work | wɛrək | ʋʏrk | ʋɛrək | ʋɛ̯rə̆k | ʋɑrə̆k |
| 3 | ship | sxɪp | skɪp | sxɪp | sxɪp | sxɪp |
| 4 | finger | vɪŋər | fɪŋər | vɪŋər | vɪŋəř | vɪŋəř |
| 5 | wine | ʋæ̯ˑĭn | ʋin | ʋɛin | ʋiˑn | ʋin |
| 6 | house | hœ̯ˑz | huˑz | hœy̆s | ys | hys |
| 7 | milk | mɛlə̆k | mɔˑlkə | mɛlək | mɛlə̆k | mɛlə̆k |

# With LED-A can be processed ...

- vowels, pulmonic consonants and voiced implosives;
- primary stress, secondary stress;

  by processing ˈ and ˌ as segments
- extra short, normal, half long, long;

  by preprocessing: ă → a, a → aa, aˑ → aaa, aː → aaaa
- aspirated, labialized, palatalized, velarized, pharyngealized, nasalized;

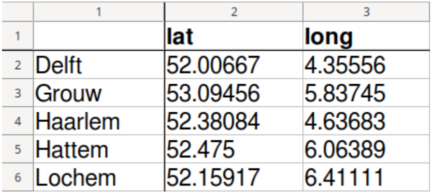
  by averaging with h, w, j, ɣ, ʕ, n respectively.

# Maps

- Easy to create maps, only coordinates are required (no outline required).

# Coordinates

Rows are items, columns are locations



|   | 1 | 2 | 3 |
|---|---|---|---|
| 1 |  | **lat** | **long** |
| 2 | Delft | 52.00667 | 4.35556 |
| 3 | Grouw | 53.09456 | 5.83745 |
| 4 | Haarlem | 52.38084 | 4.63683 |
| 5 | Hattem | 52.475 | 6.06389 |
| 6 | Lochem | 52.15917 | 6.41111 |

Coordinates are taken from GeoNames.

# Output

- Output: individual word pair distances and/or aggregated distances.

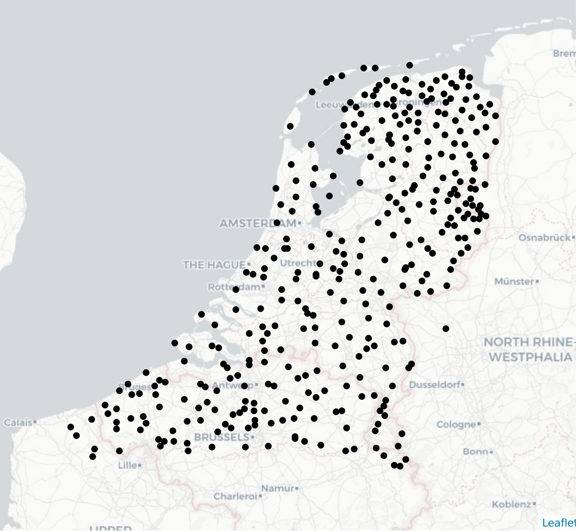
# Visualization

Example data set:

- Reeks Nederlandse Dialectatlassen, compiled by E. Blancquaert and W. Pée.
- Texts from 1922–1975, 1956 local dialects, 139 sentences each.
- We selected 360 dialects, 166 words.
- Standard Dutch and Standard German were added.

# Measurements

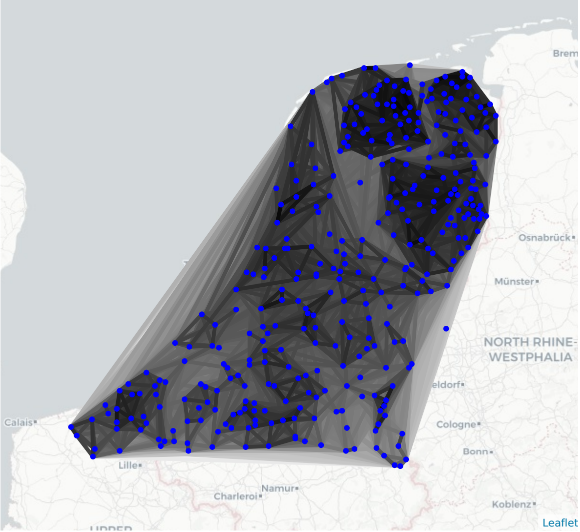
- We measure distances using PMI-based Levensthein distance.
- Length and diacritics are processed.



Distribution of the 360 varieties in the Dutch dialect area.

# Beam maps

- Introduced by Goebl (± 1983).
- The locations of local dialects are connected to each other by straight lines in a map.
- Darker lines represent smaller distances, lighter lines represent larger distances.



Beam map showing Levenshtein distances among local dialects. Darker lines represent larger distances.

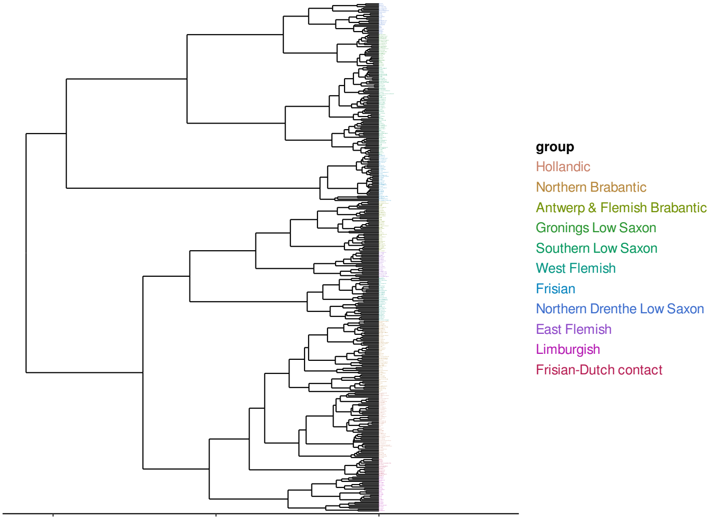
# Cluster analysis

- Introduced by Goebl (± 1982) in dialectometry.
- Group objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). (Wikipedia)

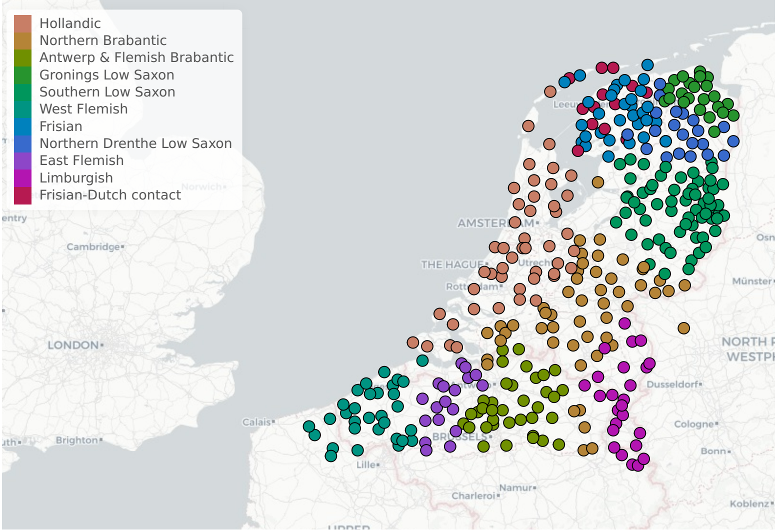
# Cluster methods

In LED-A five cluster methods are available:

- **Single-Linkage**: chaining effect, no clear group structure.
- **Complete-Linkage**: compact clusters of about equal size.
- **UPGMA**: results reflect the original distances most closely.
- **WPGMA**: in case of a more irregular distribution.
- **Ward's** method: minimizes the variance in the clusters; it usually creates compact, even-sized clusters (Szmrecsanyi, 2012)



Dendrogram obtained using Ward's method. The tree structure explains 50.8% of the variance in the original distances.



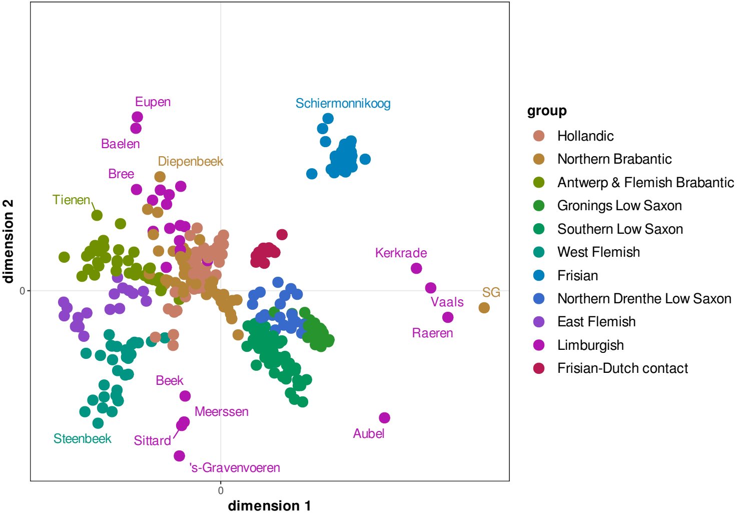
Eleven groups derived from the dendrogram that was obtained using Ward's method.

# Multidimensional scaling

- Introduced by Embleton (1993) in dialectometry.
- Put the local dialects on a map so that the distances in two-dimensional space reflect the distances in the matrix as closely as possible.

# Multidimensional scaling methods

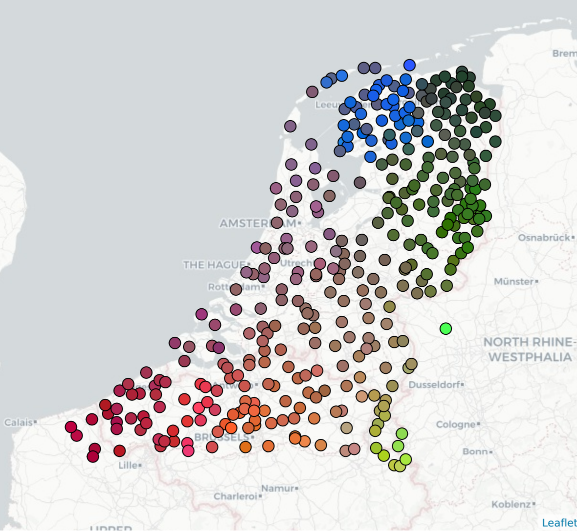
- **Classical** (metric):
  original algorithm, proposed by Togerson (1952).
- **Kruskal's** non-metric:
  results reflect the original distances most closely.
- **Sammon's** non-linear (metric) mapping:
  points are shown more dispersed.
- **t-SNE** (t-distributed stochastic neighbor embedding):
  reveals (otherwise hidden) patterns, is stochastic.



Using Kruskal's non-metric MDS the 362 dimensions are reduced to 2. They explain 78.5% of the variance in the original distances.

# Multidimensional scaling

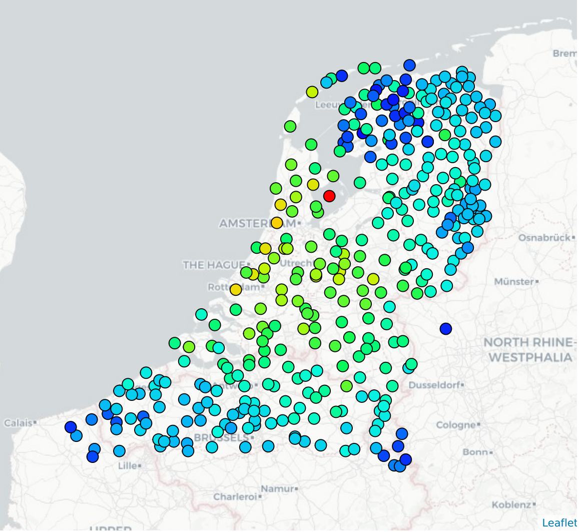
- With MDS scaling to three or more dimensions is possible as well.
- Scale to three dimensions so that each local dialect is represented by three values $x$, $y$ and $z$.
- Let $x$ be the intensity of red, $y$ be the intensity of green and $z$ be the intensity of blue.
- Introduced by Nerbonne, Heeringa & Kleiweg (1999).



RGB map where *x* determines inversely the intensity of red, *y* determines the intensity of blue and *z* determines the intensity of green.

# Reference point maps

- Introduced by Goebl (± 1982).
- Compare local dialects to a reference point (e.g. standard language, proto-language).
- Coloring according to rainbow scheme: red is most similar, blue is most distant.

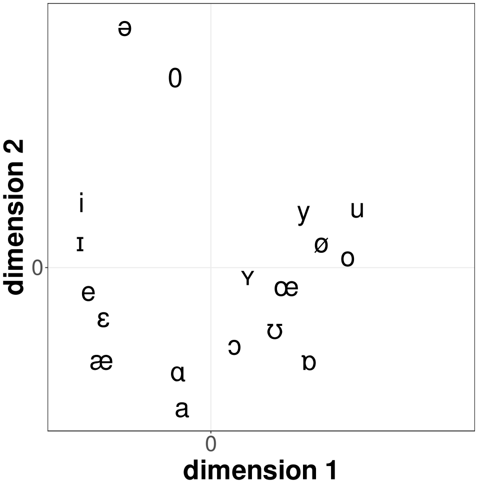


Dutch dialects compared to Standard Dutch. Red dots represent strongly related dialects, blue dots more remote ones.

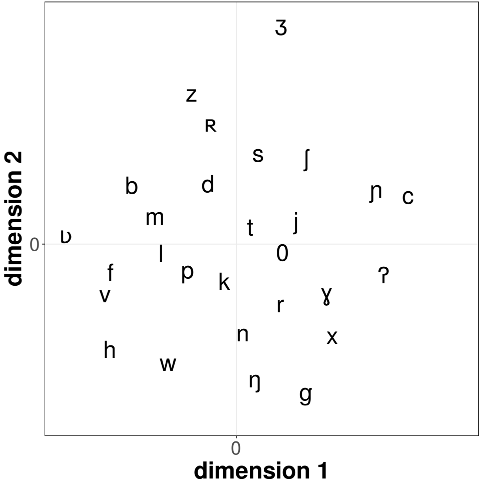
# Segment distances

- Show feature-based or PMI-based segment distances.
- Reduce the distances among the segments to two dimensions with Kruskal's non-metric MDS.

## Vowels



# Consonants


ʒ
z
ʀ
s
ʃ
b
d
ɲ
c
m
t
j
ʋ
0
l
0
f
p
ʔ
k
v
ɣ
r
n
x
h
w
ŋ
g
0
dimension 1
dimension 2

# Thanks!

# Thanks!

The slides are available at:
`led-a.org/slides.pdf`