

## Data, data documentation and analysis scripts for

*Measuring foreign accent strength in English. Validating the Levenshtein distance as a measure.*

Martijn Wieling<sup>(1,2)</sup> & Jelke Bloem<sup>(3)</sup> & Kaitlin Mignella<sup>(1)</sup> & Mona Timmermeister<sup>(4)</sup> & John Nerbonne<sup>(1,5)</sup>

<sup>1</sup>University of Groningen, the Netherlands & <sup>2</sup>Eberhard Karls University, Germany & <sup>3</sup>University of Amsterdam, the Netherlands & <sup>4</sup>Utrecht University, the Netherlands & <sup>5</sup>Freiburg Institute for Advanced Studies, Germany

Journal: **Language Dynamics and Change** (accepted for publication)

Preprint: <http://www.martijnwieling.nl/files/WielingBloemMignellaEtAl-LDC.pdf>

### Abstract

With an eye toward measuring the strengths of foreign accents in American English, we evaluate the suitability of a modified version of the Levenshtein distance (LD) for comparing accented pronunciations. Although this measure has been used successfully *inter alia* to study the differences among dialect pronunciations, it has not been applied to study foreign accents. Here, we use it to compare the pronunciation of non-native English speakers to native American English speech. Our results indicate that the Levenshtein distance is a valid native-likeness measurement, as it correlates strongly with the average “native-like” judgments given by more than 1000 native American English raters ( $r = -0.8$ ,  $p < 0.001$ ).

**Keywords:** foreign accent, Levenshtein distance, edit distance, pronunciation, validation.

# 1 Packages and functions

```
R.Version()$version.string  
  
## [1] "R version 3.1.0 (2014-04-10)"  
  
# cronbach alpha function the package ltm (version 0.9.9)  
source('functions/cronbach.alpha.R')
```

## 2 English accents data sets

```
load("data/judgeAccentsLD.rda")  
load("data/accntRatings.rda")
```

Legenda judgeAccentsLD (286 observations of 4 variables):

1. Speaker : the speaker
2. Nativelikeness : the average nativelikeness rating given by native U.S. English-speaking judges
3. DistLD.nodia : average PMI-based Levenshtein distance with respect to the 115 U.S. English speakers in the data set (normalized for length, diacritics are ignored, the PMI-based sound distances are based on transcriptions from 989 speakers)
4. DistLD.asplength : average PMI-based Levenshtein distance with respect to the 115 U.S. English speakers in the data set (normalized for length, aspiration and lengthening are marked, but other diacritics are ignored, the PMI-based sound distances are based on transcriptions from 989 speakers)
5. DistLD.nodia.noFrenchSpanishInPMI : average PMI-based Levenshtein distance with respect to the 115 U.S. English speakers in the data set (normalized for length, diacritics are ignored, the PMI-based sound distances are based on transcriptions from 891 speakers – 64 Spanish speakers and 34 French speakers were excluded)

Legenda accntRatings (1143 observations of 290 variables):

1. ID : participant ID
2. Gender : the gender of the participant
3. Age : the age of the participant
4. State : the state where the participant was born
5. english23 ... korean17 : the samples for which ratings were obtained (corresponding with the Speech Accent Archive ID)

### 3 Analysis and results

```
# gender distribution of participants
table(accentRatings$Gender)

##
##      F      M
## 485 658

# state distribution of participants
table(accentRatings$State)

##
##      Alabama      Alaska      Arizona      Arkansas      California
##          12          4          7          7          151
##      Colorado      Connecticut      Delaware      Florida      Georgia
##          21          16          3          22          20
##      Hawaii      Idaho      Illinois      Indiana      Iowa
##          6          1          64          12          14
##      Kansas      Kentucky      Louisiana      Maine      Maryland
##          10          3          15          7          28
##      Massachusetts      Michigan      Minnesota      Mississippi      Missouri
##          68          38          23          3          17
##      Montana      Nebraska      New Hampshire      New Jersey      New Mexico
##          4          4          11          42          7
##      New York      North Carolina      North Dakota      Ohio      Oklahoma
##          115          19          1          66          12
##      Oregon      Pennsylvania      Rhode Island      South Carolina      South Dakota
##          20          54          5          10          5
##      Tennessee      Texas      Utah      Vermont      Virginia
##          9          55          14          7          36
##      Washington      Washington DC      West Virginia      Wisconsin      Wyoming
##          42          13          2          15          2

# average age of participants (and standard deviation)
mean(accentRatings$Age)

## [1] 36.21

sd(accentRatings$Age)

## [1] 13.87

# average number of rated samples per participant (and standard deviation)
mean(rowSums(!is.na(accentRatings[,5:ncol(accentRatings)])))

## [1] 41

sd(rowSums(!is.na(accentRatings[,5:ncol(accentRatings)])))

## [1] 14.02
```

```

# minimum number of judgements per sample
min(colSums(!is.na(accentRatings[,5:ncol(accentRatings)])))

## [1] 59

# cronbach's alpha of the ratings
cronbach.alpha(accentRatings[,c(5:ncol(accentRatings))],na.rm=T)

##
## Cronbach's alpha for the 'accentRatings[, c(5:ncol(accentRatings))]' data-set
##
## Items: 286
## Sample units: 1143
## alpha: 0.853

```

```

# native language distribution of the speakers
# note that this table includes 6 speakers who are native English speakers
# and have been born in the United States
sort(table(gsub("\\d", "", judgeAccentsLD$Speaker)),dec=T)

##
##      spanish      english      french      arabic      german      portuguese
##      17           14           13           12           9           9
##      italian     russian     japanese     korean     mandarin     dutch
##      8            8            7            7            7            6
##      hindi       turkish      farsi       greek      hungarian     romanian
##      6            6            5            5            5            5
##      bulgarian   cantonese    hausa      norwegian    thai          albanian
##      4            4            4            4            4            3
##      amharic     armenian     bengali     bosnian     finnish       gujarati
##      3            3            3            3            3            3
##      kiswahili    krio         polish      ukrainian   afrikaans     bambara
##      3            3            3            3            2            2
##      croatian    hadiyya      hebrew      indonesian   kambaata      khmer
##      2            2            2            2            2            2
##      kurdish     marathi      punjabi     slovenian    somali         swedish
##      2            2            2            2            2            2
##      taiwanese   tibetan      uzbek       vietnamese   yoruba        akan
##      2            2            2            2            2            1
##      azerbaijani  бага         bai         bamun       basque        catalan
##      1            1            1            1            1            1
##      cebuano      czech        danish      edo         fang           ga
##      1            1            1            1            1            1
##      georgian     gusii        igbo        kanuri      lamotrekese    lao
##      1            1            1            1            1            1
##      latvian     lithuanian    luo         macedonian   mandingo       mandinka
##      1            1            1            1            1            1
##      mende       mongolian    moore       oriya       pashto        pulaar
##      1            1            1            1            1            1
##      quechua     serbian      sesotho     sindhi      sinhalese     swissgerman

```

```
##          1          1          1          1          1          1
##    tagalog      tamil      telugu  tetundili  tokpisin    tswana
##          1          1          1          1          1          1
##    vlaams      yupik      zulu
##          1          1          1
```

```
# total number of unique languages
length(unique(gsub("\\d","", judgeAccentsLD$Speaker)))
```

```
## [1] 99
```

```
# languages spoken by a single speaker in this dataset
sum(table(gsub("\\d","", judgeAccentsLD$Speaker))==1)
```

```
## [1] 46
```

```
# native languages distribution of the data used to determine PMI distances
all = read.table('data/generate-LD-distances/allprons.csv',header=T,sep='\t')
colnames(all)[1] = 'Speaker'
```

```
# number of native American English speakers from the United States
length(strsplit(as.character(all[all$Speaker=='United States'],$please),' / ')[[1]])
```

```
## [1] 115
```

```
all = all[all$Speaker != 'United States',] # excluding the United States reference
sort(table(gsub("\\d","", all$Speaker)),dec=T)
```

```
##
##    english      spanish      french      arabic
##      181         64         34         31
##    portuguese    russian    mandarin    german
##      27         26         25         22
##    italian      korean      turkish    cantonese
##      21         21         20         17
##    romanian     dutch      serbian     amharic
##      17         15         14         13
##    polish       farsi      bengali     bulgarian
##      13         12         11         11
##    japanese     swedish     hindi     vietnamese
##      11         10         9         9
##    bosnian      hausa      hebrew     kiswahili
##      8          8          8          8
##    thai         greek      hungarian  urdu
##      8          7          7          7
##    albanian     croatian    gujarati   tagalog
##      6          6          6          6
##    taiwanese    ukrainian    armenian   bambara
##      6          6          5          5
##    czech        georgian    indonesian krio
```

##	5	5	5	5
##	kurdish	mongolian	norwegian	wolof
##	5	5	5	5
##	finnish	khmer	lithuanian	macedonian
##	4	4	4	4
##	nepali	pashto	punjabi	sinhalese
##	4	4	4	4
##	tigrigna	yoruba	afrikaans	dari
##	4	4	3	3
##	fijian	icelandic	igbo	kambaata
##	3	3	3	3
##	marathi	pulaar	slovak	swissgerman
##	3	3	3	3
##	tamil	tibetan	uzbek	azerbaijani
##	3	3	3	2
##	bafang	bari	basque	bavarian
##	2	2	2	2
##	belarusan	catalan	danish	ewe
##	2	2	2	2
##	ga	hadiyya	kikongo	lao
##	2	2	2	2
##	latvian	luo	mende	oriya
##	2	2	2	2
##	oromo	quechua	rotuman	satawalese
##	2	2	2	2
##	slovenian	somali	telugu	tswana
##	2	2	2	2
##	twi	uyghur	yiddish	agni
##	2	2	2	1
##	akan	amazigh	baga	bai
##	1	1	1	1
##	balantaganja	bamun	burmese	carolinian
##	1	1	1	1
##	cebuano	chamorro	chichewa	chittagonian
##	1	1	1	1
##	dinka	ebira	edo	estonian
##	1	1	1	1
##	fang	fanti	fataluku	frisian
##	1	1	1	1
##	fulfuldeadamawa	ganda	gusii	hainanese
##	1	1	1	1
##	ibibio	jola	kannada	kanuri
##	1	1	1	1
##	kazakh	kikuyu	kirghiz	konkani
##	1	1	1	1
##	lamaholot	lamotrekese	malay	malayalam
##	1	1	1	1
##	maltese	mandingo	mandinka	mankanya
##	1	1	1	1
##	mauritian	moba	moore	mortlockese
##	1	1	1	1

```
##          nama          nandi          newari          patois
##          1            1            1            1
##    pohnpeian      poonchi      rwanda      sardinian
##          1            1            1            1
##          sarua      serer      sesotho      shilluk
##          1            1            1            1
##          shona      sicilian      sindhi      sundanese
##          1            1            1            1
##          susu      sylheti      taishan      tatar
##          1            1            1            1
##    teochew      tetundili      tokpisin      vlaams
##          1            1            1            1
##    xasonga      yapese      yupik      zulu
##          1            1            1            1
```

```
# total number of unique languages
length(unique(gsub("\\d","", all$Speaker)))
```

```
## [1] 172
```

```
# correlations when ignoring all diacritics
cor( judgeAccentsLD$Nativelikeness, judgeAccentsLD$DistLD.nodia )
```

```
## [1] -0.7707
```

```
cor( judgeAccentsLD$Nativelikeness, log(judgeAccentsLD$DistLD.nodia) )
```

```
## [1] -0.8102
```

```
# correlations when aspiration and lengthening are distinguished
cor( judgeAccentsLD$Nativelikeness, judgeAccentsLD$DistLD.asplength )
```

```
## [1] -0.779
```

```
cor( judgeAccentsLD$Nativelikeness, log(judgeAccentsLD$DistLD.asplength) )
```

```
## [1] -0.8147
```

```
# the effect of the languages included in the PMI measure is limited
options(digits=6)
cor( judgeAccentsLD$DistLD.nodia,
      judgeAccentsLD$DistLD.nodia.noFrenchSpanishInPMI, use="pairwise.complete.obs" )
```

```
## [1] 0.999984
```

```
# average agreement of the individual raters with mean ratings
# the ratings of the individual raters are excluded from the correlation
scores = accentRatings[,c(5:ncol(accentRatings))] # select ratings
summedRatings = as.data.frame(colSums(scores,na.rm=T)) # sum of all ratings
nrRatings = as.data.frame(colSums(scores >=1 ,na.rm=T)) # number of ratings
```



```

cnt = 0
cors = 0
ratings = t(scores)
for (i in 1:ncol(ratings)) {
  correctedRatings = (summedRatings - ratings[,i]) / (nrRatings - 1)
  if (sum(!is.na(correctedRatings)) > 1) {
    cors = cors + cor(correctedRatings, ratings[,i], use='pairwise.complete.obs')
    cnt = cnt + 1
  }
}

avgcor = cors / ncol(ratings)
rownames(avgcor) = NULL
avgcor

##           [,1]
## [1,] 0.838608

# plotting the relationship between the Levenshtein distance
# and average native-likeness ratings
plot(log(judgeAccentsLD$DistLD.nodia), judgeAccentsLD$Nativelikeness,
     xlab='PMI-based Levenshtein distance (log)',
     ylab='Average native-likeness ratings')

# adding regression line to plot
abline(lm(Nativelikeness ~ log(DistLD.nodia), data=judgeAccentsLD))

```

