

# Regression Analysis on Levenshtein- Pointwise Mutual Information

## Segment Distances Across Languages and Acoustic Distances

Eliza Margaretha  
Martijn Wieling  
John Nerbonne

Rijksuniversiteit Groningen

---

# Outline

- Overview
- Techniques
- Data
- Analysis
- Discussion
- Summary

# Overview

- Compare phonetic segment distances
  - Dutch, German, Bulgarian
- Compare Levenshtein-Pointwise Mutual Information (PMI) distances to acoustic distances
- Regression analysis
  - Correlation
  - Prediction power

# Techniques: Levenshtein-PMI (1/3)

- Segment Distance
  - How often segment  $x$  is aligned with segment  $y$
- Levenshtein
  - Insertion: a segment with a gap
  - Deletion: a gap with a segment
  - Substitution: 2 segment

# Techniques: Levenshtein-PMI (2/3)

- Pointwise Mutual Information (Church and Hanks, 1995)

$$PMI(x,y) = \log_2 \left( \frac{p(x,y)}{p(x)p(y)} \right)$$

- Wieling, et al. (2009)
  - $p(x,y)$  is the number of the  $x$  and  $y$  occurrences at the same position in 2 aligned strings of  $X$  and  $y$ , divided by the total number of aligned segments
  - $p(x)$  or  $p(y)$  the number of the occurrences of  $x$  or  $y$  divided by the total number of segment occurrences

# Techniques: Levenshtein-PMI (2/3)

- Training Wieling, et al. (2009)
  1. Align string with Levenshtein algorithm (w/o vocal-consonant)
  2. Compute PMI values and transform (subtract from 0 + max value)
  3. Apply Levenshtein to PMI-segment distances
  4. Repeat 2 and 3 till convergence is reached

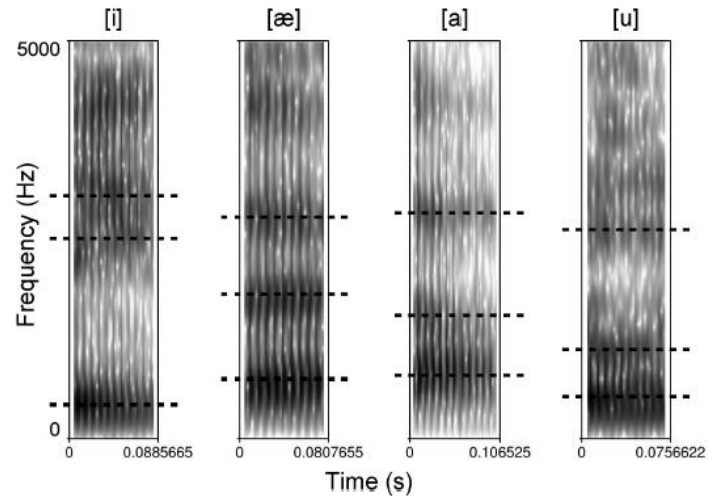
	ɑ	ə	ɛ
ɑ	0	2331	1880
ə	2331	0	64905
ɛ	1880	64905	0

# Techniques: Formant Measurements (1/3)

- ③ Vowel quality (McArthur, 1998)
  - the property that makes vowels different, e.g. /i:/ as in sheep from /ɪ/ as in ship
  - determined by the position of the vocal tracts during pronunciation
- Formants
  - measure vowel quality by means of acoustic signals
  - specify the energy concentration positions in the acoustic signals, i.e. the lowest resonance frequencies (Peterson & Barney, 1952)

# Techniques: Formant Measurements (2/3)

- ③ Formants: darker bands
  - 2 first formants are the most distinguishing
  - 3<sup>rd</sup> formants and lip position
- */i/* and */u/* has similar first formants but the second formant of */i/* is much higher than that of */u/*



Picture from (Leinonen, 2010)

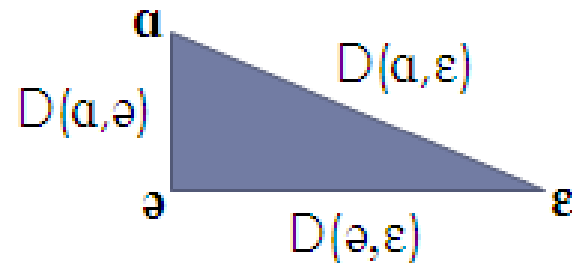


# Techniques: Formant Measurements (3/3)

- Acquiring acoustic distances
  - Compute Euclidean distances of formant values between vowel pairs (Wieling, et al., 2007)
- Normalizing non-linguistic speaker-dependent differences
  - Pitch, gender, shape & size of vocal tracts
  - transforms Hertz frequency to the Bark or the Mel scales

# Techniques: Mantel Test (1/2)

- Triangle inequality
  - Dependent:  $D(a, \epsilon)$  is dependent to  $D(a, \theta)$  and  $D(\theta, \epsilon)$ 
    - $D(a, \epsilon) < D(a, \theta) + D(\theta, \epsilon)$
    - Acoustic distance
  - Independent
    - Levenshtein PMI
- Mantel test
  - Significance Test of a Correlation Coefficient of Distance Matrices



# Techniques: Mantel Test (2/2)

- Random permutation test
- H Null = No relation between 2 matrices
  - R should be equally likely to be larger or smaller
- Steps
  1. Permutate rows and columns of one of the matrices randomly
  2. Compute correlation between the permuted matrix and the other matrix
  3. Repeat 1 and 2
- Observation value:
  - Add 1 for every  $r(D1, D2) > r(D1, D2)$
  - Divided by number of repetition

# Dataset (1/5)

Language	Locations	Words	Segment Types
Dutch	424	562	82
German	186	196	78
Bulgarian	197	152	67

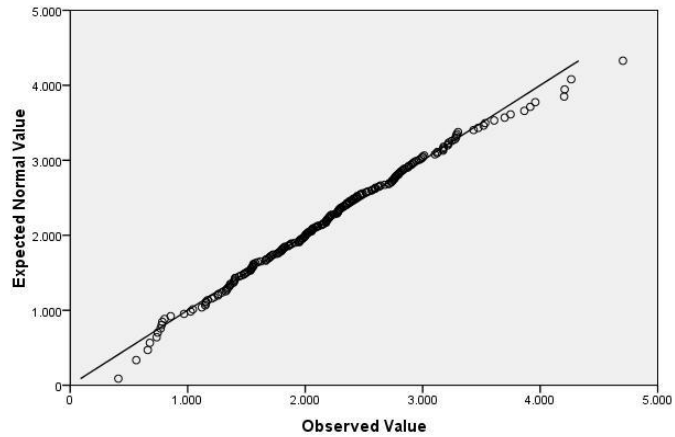
- **Dutch:** Goeman-Taeldean-Van Reenen-Project
- **German :** Kleiner Deutscher Lautatlas project
- **Bulgarian:** students' theses at the University of Sofia, published monographs, dictionaries, and the archive of the Ideographic Dictionary of Bulgarian Dialects (Prokić, et al., 2009)

# Dataset (2/5)

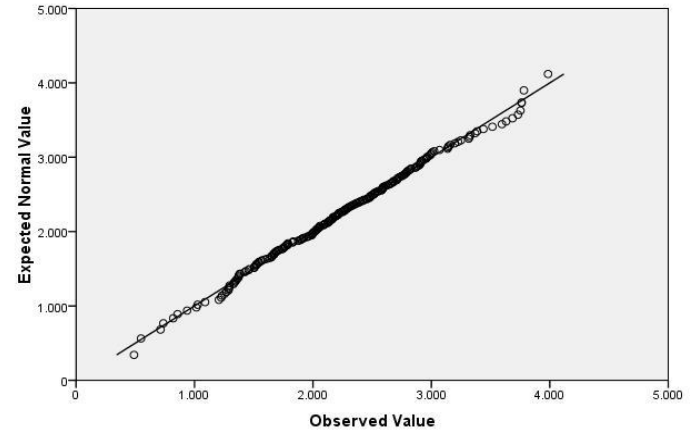
Language Pair	Shared Types	Segment Alignments	Vowel Alignments	Consonant Alignments
Dutch and Bulgarian	43	235	92	143
Dutch and German	71	870	261	609

# Dataset (3/5)

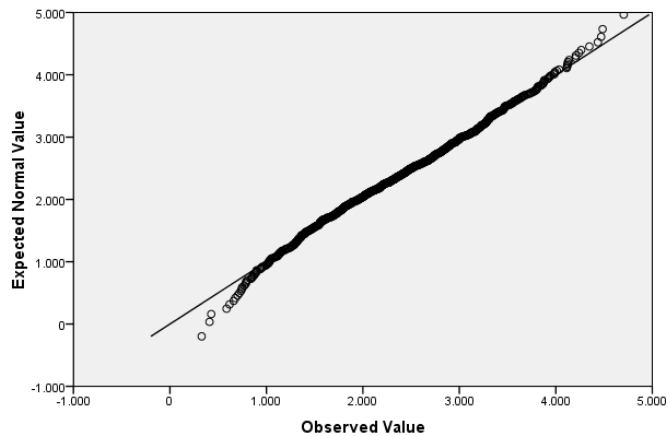
Normal Q-Q Plot of NL



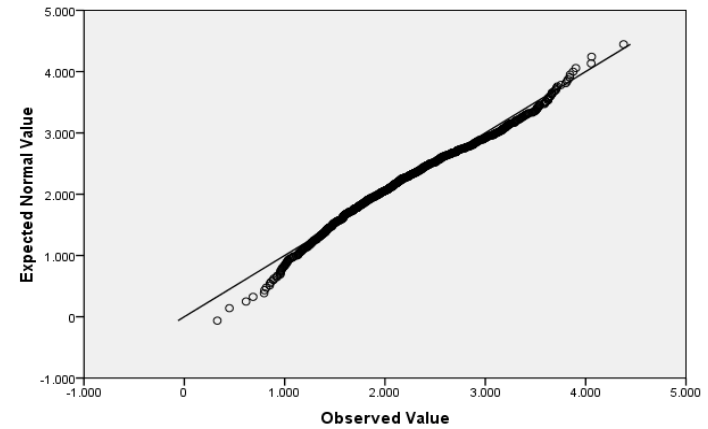
Normal Q-Q Plot of BUL



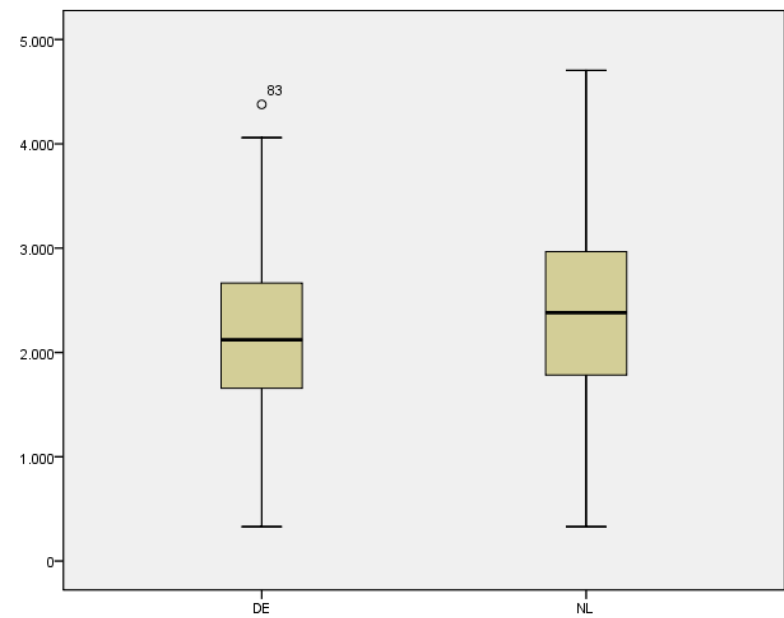
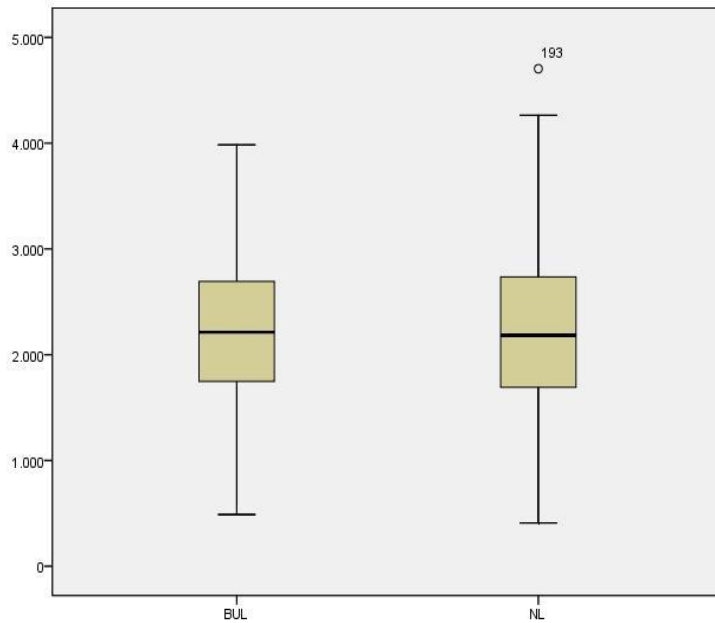
Normal Q-Q Plot of NL



Normal Q-Q Plot of DE



# Dataset (4/5)



## Dataset (5/5)

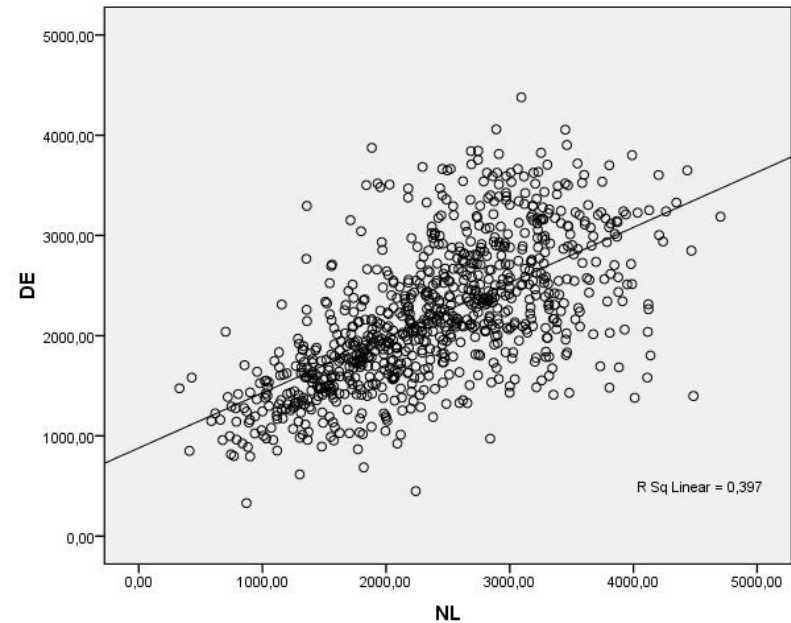
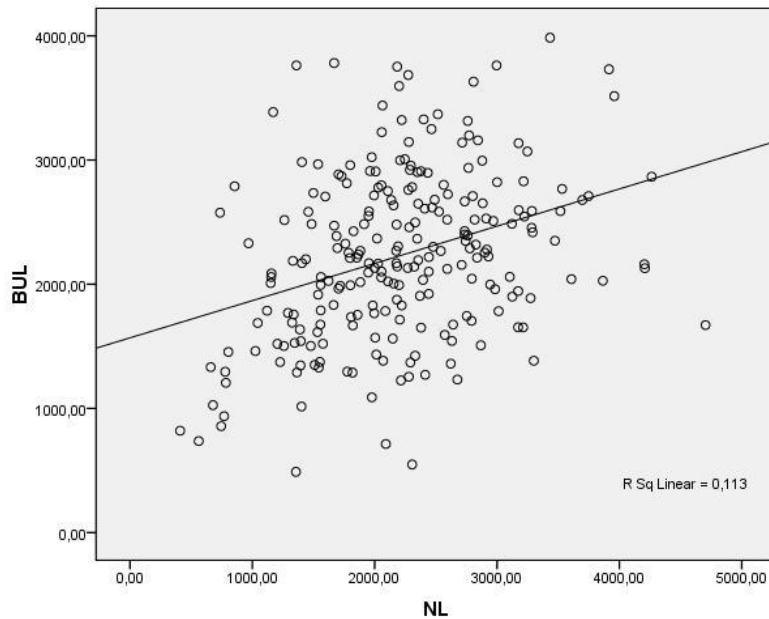
- Acoustic data was obtained from Pols, et al. (1973) and Van Nierop, et al. (1973),
  - three first formants
  - 50 male and 25 female Dutch speakers
  - 36 acoustic vowel alignments
  - All alignments appear in Levenshthein-PMI Dutch matrix



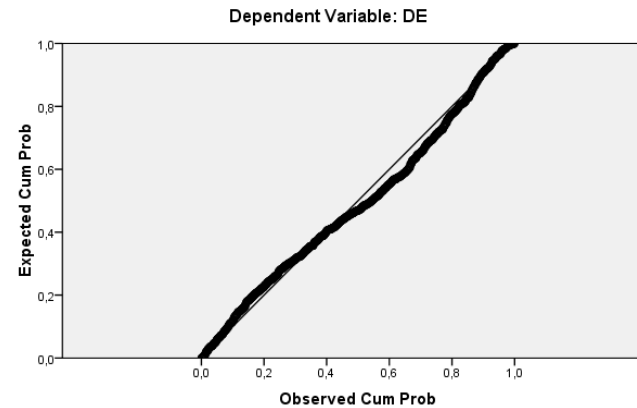
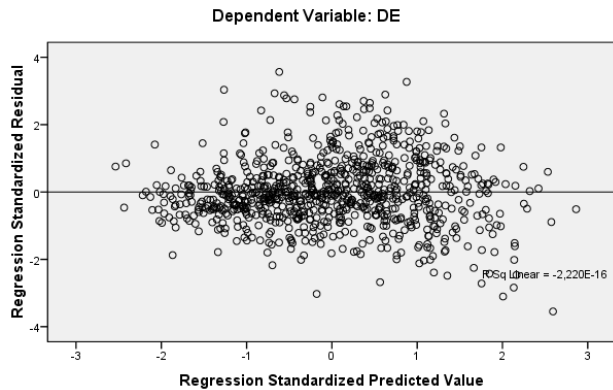
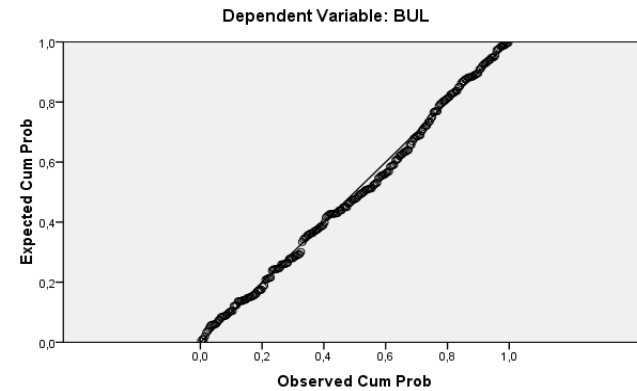
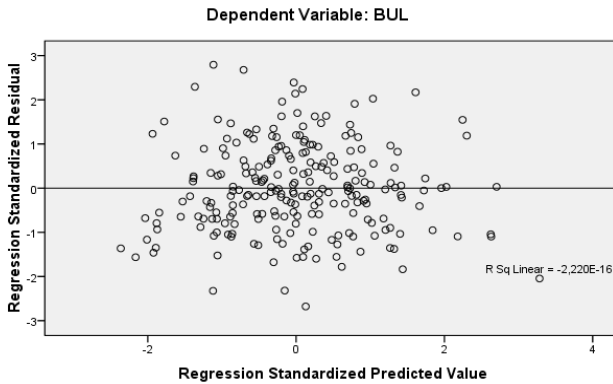
# Analysis: Lev-PMI Distance Across Languages (1/5)

- Regression analysis setup
  - Variables
    - Dutch (independent/explanatory) and Bulgarian (dependent/response)
    - Dutch and German
  - Cases
    - Segment alignments
  - Values
    - Levenshtein-PMI distance

# Analysis: Lev-PMI Across Languages (2/5)



# Analysis: Lev-PMI Across Languages (3/5)



# Analysis: Lev-PMI Across Languages (4/5)

Language Pair	Alignment Sets	Pearson Correlation	Effect size ( $r^2$ )
Dutch and Bulgarian	All	0,336	0.113
	Vowel	0,418	0.178
	Consonant	0,339	0.115
Dutch and German	All	0,630	0,397
	Vowel	0,620	0.384
	Consonant	0.587	0.345

# Analysis: Lev-PMI Across Languages (5/5)

- Computing regression line
  - $y = 1586,562 + 0,300x$
  - T ratio 5,454 ( $p < 0,001$ )

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1568,562	128,322		12,224	,000
	NL	,300	,055	,336	5,454	,000

a. Dependent Variable: BUL

# Analysis: Example (1/2)

- How does the prediction work?
- Lev-PMI distance between **a** and  $\varepsilon$  in Dutch,  $x = 1556$
- Predicted **a**- $\varepsilon$  distance in Bulgarian:
  - $\hat{y} = 1586.562 + 0.300 (1556) = 2053.362$

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,336 <sup>a</sup>	,113	,109	639,39919

a. Predictors: (Constant), NL

b. Dependent Variable: BUL

**Descriptive Statistics**

	Mean	Std. Deviation	N
BUL	2230,4298	677,53592	235
NL	2207,7872	760,44125	235

## Analysis: Example (2/2)

$$SE_{\hat{y}} = s \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i^n (x_i - \bar{x})^2}}$$

- $SE_{\hat{y}} = 639.4 \times \sqrt{\frac{1}{235} + \frac{(1556 - 2207.8)^2}{578270.9}} = 549.6$
- $t$  for (df = 200) = 1.97 ( $\alpha = 0.05$ )
- *Confidence Interval* 95% =  $\hat{y} \pm t \times SE_{\hat{y}} = 2053.362 \pm 1.97 \times 549.6 = 2053.362 \pm 1083$
- With 95% certainty, mean of  $\alpha$ - $\varepsilon$  distance in Bulgarian given the distance in Dutch = 1556, lies in the interval (970,3136).
- Real distance = 1675

# Analysis: Lev-PMI and Acoustic Distances (1/3)

- Response Variable
  - Lev-PMI distance for Dutch segments
- Explanatory variables (acoustic distance variations)
  - Hertz: raw hertz measurements of formants
  - Bark: hertz values transformed to Bark scale
  - Mel: hertz values transformed to Mel scale
  - Z-score
    - hertz values transformed to Z-scores per speaker, normalizing over all the vowels for each speaker
    - average the Z-scores per vowel of all speakers



# Analysis: Lev-PMI and Acoustic Distances (2/3)

Acoustic variation	Number of first formants	Pearson Correlation	Effect Size ( $r^2$ )	Significance
Hertz	2	0.481	23 %	0.003
	3	0.426	18 %	0.010
Z-score	2	0.720	<b>52 %</b>	0.000
	3	0.640	<b>41 %</b>	0.000
Bark Scale	2	0.616	38 %	0.000
	3	0.517	27 %	0.001
Mel Scale	2	0.603	36 %	0.000
	3	0.507	26 %	0.002

# Analysis: Lev-PMI and Acoustic Distances (3/3)

- Mantel test with 9999 replicates
- $H_0$  = No relation between Lev-PMI distance with Acoustic distance
- Positive observations shows positive relationships

Acoustic variation	Observation value	Significance (p-value)
Hertz 2	0.168	0.0134
Hertz 3	0.132	<b>0.035</b>
Z2	0.410	1e-04
Z3	0.317	3e-04
Bark 2	0.303	2e-04
Bark 3	0.206	0.0027
Mel 2	0.286	2e-04
Mel 3	0.195	0.0036

# Discussion

- Why is the correlation between Dutch and Bulgarian smaller than that between Dutch and German?
- Why do Z-scores yield better results than other variations (Hertz, Bark, Mel)?
- How are the relationships between Levenshtein-PMI distances and acoustic distances of other languages?

# Summary

- Our results show that Levenshtein-PMI distances of Dutch are able to predict those of Bulgarian and German.
- Prediction of languages in the same category / with similar characteristics (Dutch-German) is better than those with different characteristics (Dutch-Bulgarian).
- Vowel quality as represented by acoustic distances correlate reasonably highly with Levenshtein-PMI distances, particularly in our Dutch case, the former can predict up to 52% of the latter.