Questionnaire Design and Regression: English Accents Project

Mona Timmermeister & Kaitlin Mignella

Overview

- Background and experimental design
- Regression results
- Correlation results
- Questionnaire design

Foreign Accents in English

- Levenshtein measurement technique
  - Measurement of phonetic distance between two sequences of sounds
  - Calculated by counting the number of deletions, insertions and substitutions required to transform one string into the other

Main Research Question

- Can we validate this technique?
  - How well does the Levenshtein measurement predict native English speakers' judgments about how native-like a particular accent is?

Speech Samples

- The Speech Accent Archive (http://accent/gmu/edu)
- Samples of speakers reading aloud the same elicitation passage:
  - *Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.*
- 50 samples selected (one male and one female for 25 different languages)
  - In the U.S for <1 year
  - Between ages of 18 and 35

Main Research Question

- How well does the Levenshtein measurement predict native English speakers' judgments about how native-like a particular accent is?

- Calculate the Levenshtein distance between each sample and a standard American English sample
- Have others judge how native-like a particular speaker sounds

Questionnaire

- Part I – questions about the participant
- Part II – participant listens to 10 randomly selected samples
  - How native-like is this speaker?
  - How good is this speaker's pronunciation?
  - How strong of an accent does this speaker have?
  - How easy to understand is this speaker?
  - How different from your own accent?
- Part III – memory task
  - Short clips from 10 samples
  - Participant must decide if they have heard each speaker before

Data

	NameLang	Language	Native	Comprehensible	StrongAcc	Pronunciation	Levenshtein
94	Spanish	10	6	6	2	5	0,1786510
95	Spanish	10	7	7	1	7	0,1786510
96	Spanish	10	7	7	1	7	0,1786510
97	Italian	11	1	5	7	4	0,2249620
98	Italian	11	2	4	5	3	0,2249620
99	Italian	11	2	4	7	4	0,2249620
100	Italian	11	4	5	4	4	0,2249620
101	Italian	11	6	5	6	5	0,2249620
102	Italian	12	2	3	6	3	0,1685030
103	Italian	12	2	4	5	4	0,1685030
104	Italian	12	3	5	5	6	0,1685030
105	Italian	12	3	5	6	3	0,1685030
106	Italian	12	3	7	3	5	0,1685030
107	Italian	12	4	5	4	4	0,1685030
108	Italian	12	4	7	3	5	0,1685030
109	Italian	12	5	6	2	6	0,1685030
110	Italian	12	5	6	3	5	0,1685030
111	Mandarin	13	1	3	7	3	0,1894420
112	Mandarin	13	1	4	7	3	0,1894420
113	Mandarin	13	1	4	7	4	0,1894420
114	Mandarin	13	1	5	5	5	0,1894420
115	Mandarin	13	2	4	5	5	0,1894420
116	Mandarin	13	3	4	3	4	0,1894420
117	Mandarin	13	3	4	5	4	0,1894420
118	Mandarin	13	5	6	4	4	0,1894420
119	Mandarin	13	7	7	2	7	0,1894420
120	Mandarin	14	1	2	7	2	0,1230480

Regression

- How well does the Levenshtein distance predict native speakers judgments about how native-like a particular accent is?

Regression Analysis

Correlations

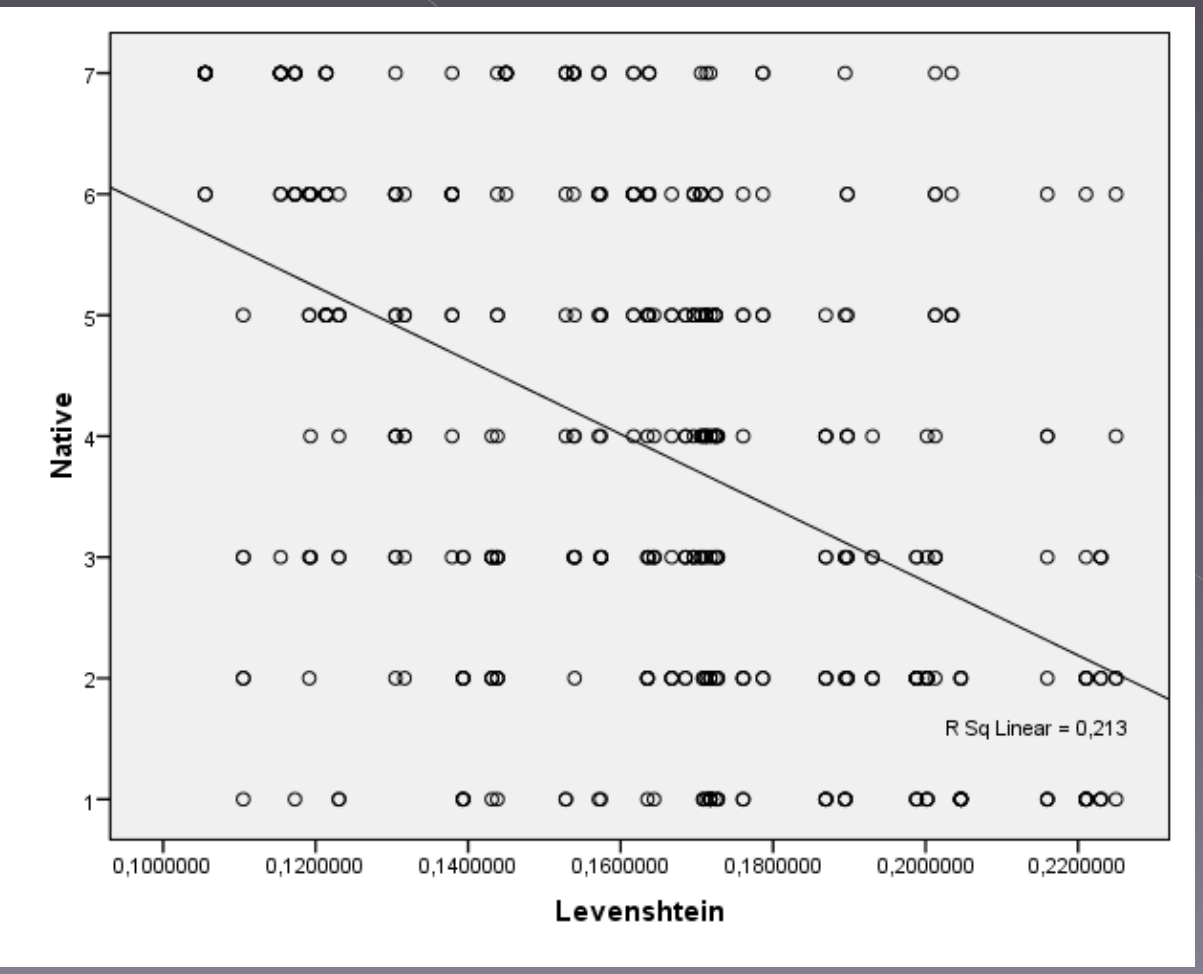
		Native	Levenshtein
Pearson Correlation	Native	1,000	-,461
	Levenshtein	-,461	1,000
Sig. (1-tailed)	Native	.	,000
	Levenshtein	,000	.
N	Native	463	463
	Levenshtein	463	463

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8,893	,455		19,528	,000
	Levenshtein	-30,470	2,731	-,461	-11,159	,000

a. Dependent Variable: Native

Regression Analysis II



R^2= 0,213

Slope = -30,470

Intercept = 8,893

[Judgment of how native-like a speaker is] = -30,470

*[Levenshtein Distance] + 8,893

Problems

- Judgments are mostly from non-native speakers
- Should we remove the data for the judgments about native English accents?
- Which “standard” U.S. English speaker should the samples be compared to?

Correlations

- Do the two questions address the same aspect?

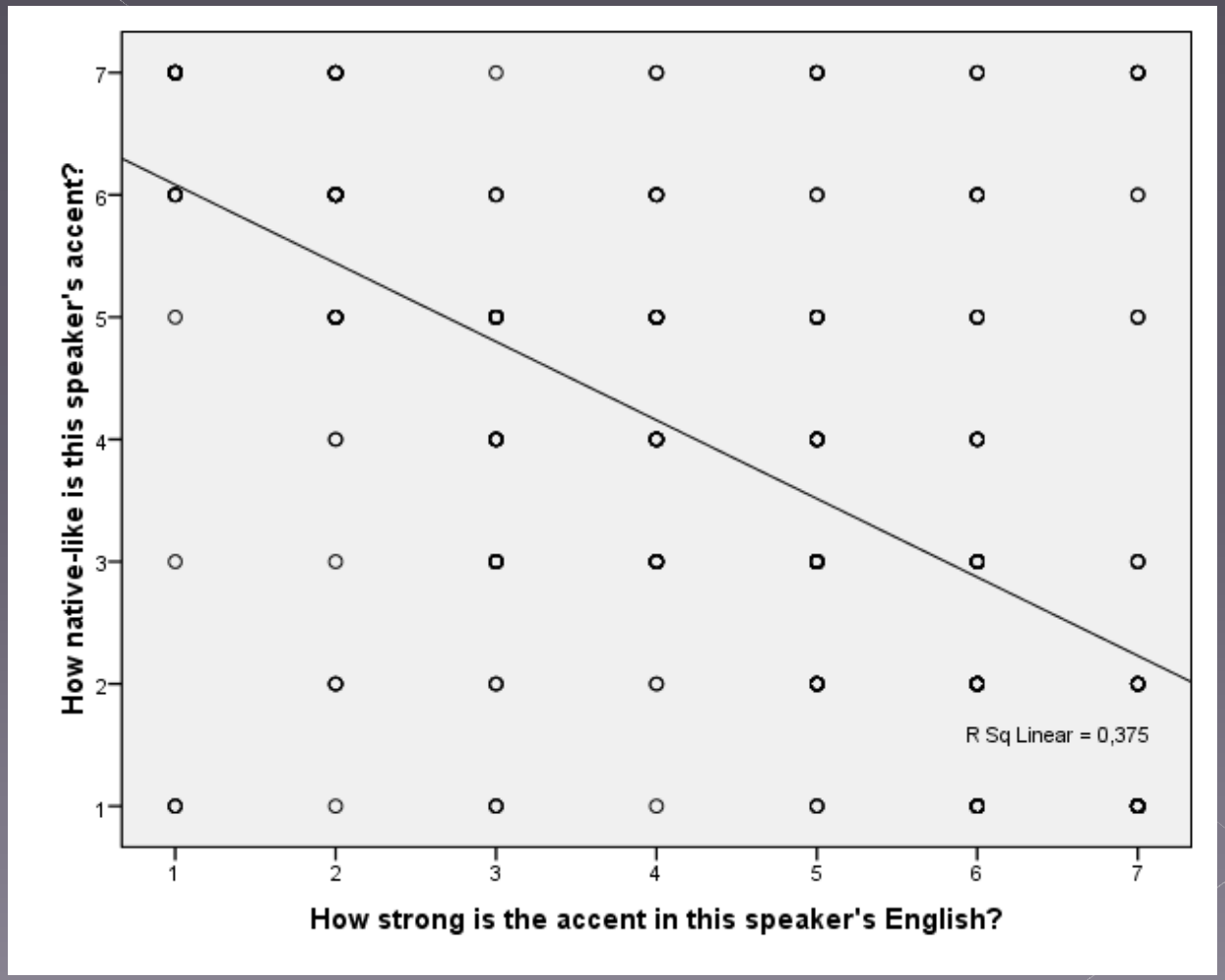
Correlations

			How native-like is this speaker's accent?	How strong is the accent in this speaker's English?
Spearman's rho	How native-like is this speaker's accent?	Correlation Coefficient	1,000	-,611**
		Sig. (2-tailed)	.	,000
		N	464	463
	How strong is the accent in this speaker's English?	Correlation Coefficient	-,611**	1,000
		Sig. (2-tailed)	,000	.
		N	463	463

**. Correlation is significant at the 0.01 level (2-tailed).

- Check if partcipants pay attention to the questions

Graph Correlation 1



Correlation 2

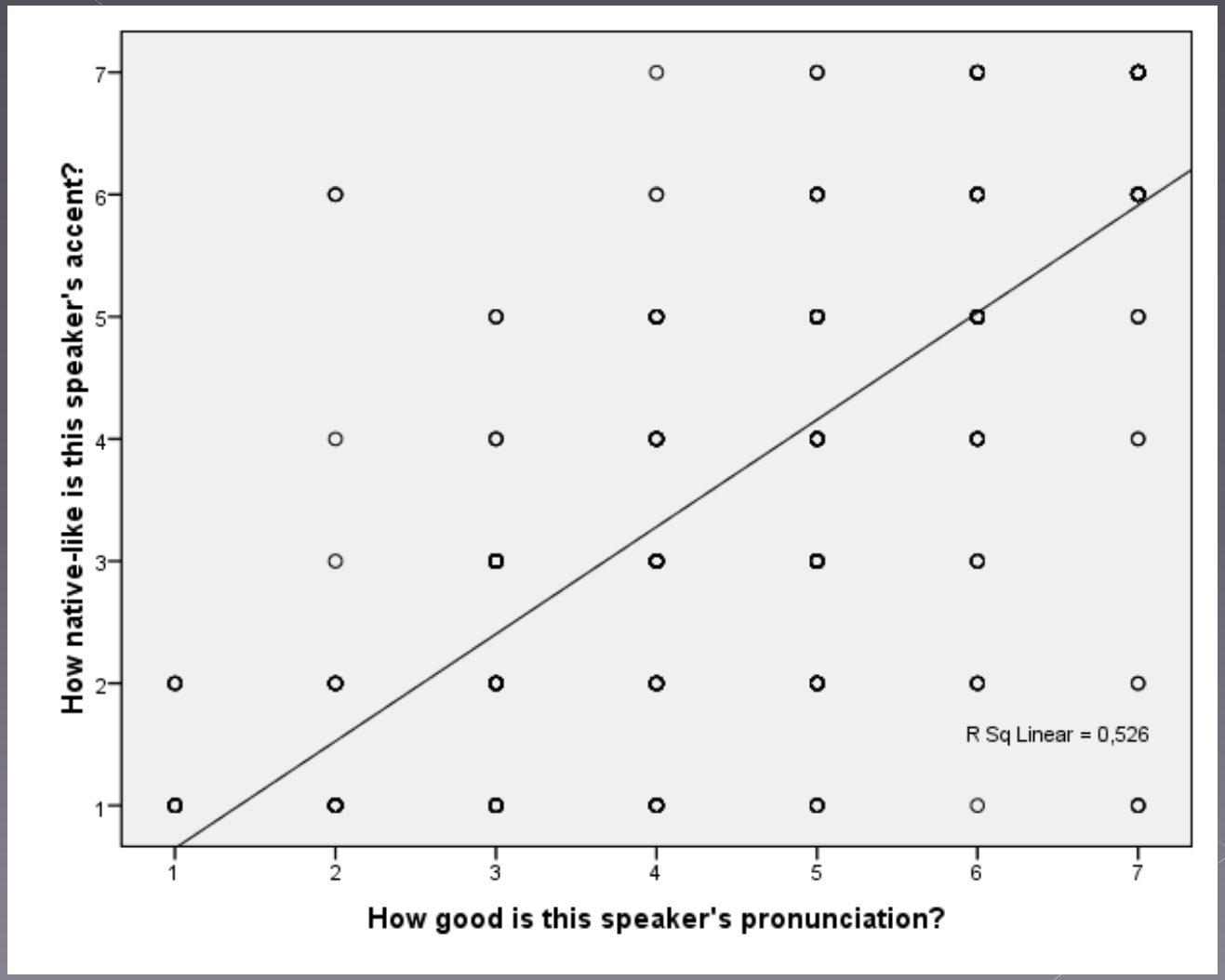
- Is there a correlation between how native-like a speaker is and how good his/her pronunciation is?

Correlations

			How native-like is this speaker's accent?	How good is this speaker's pronunciation ?
Spearman's rho	How native-like is this speaker's accent?	Correlation Coefficient	1,000	,728**
		Sig. (2-tailed)	.	,000
		N	464	463
	How good is this speaker's pronunciation?	Correlation Coefficient	,728**	1,000
		Sig. (2-tailed)	,000	.
		N	463	463

**. Correlation is significant at the 0.01 level (2-tailed).

Graph Correlation 2



Correlation 3

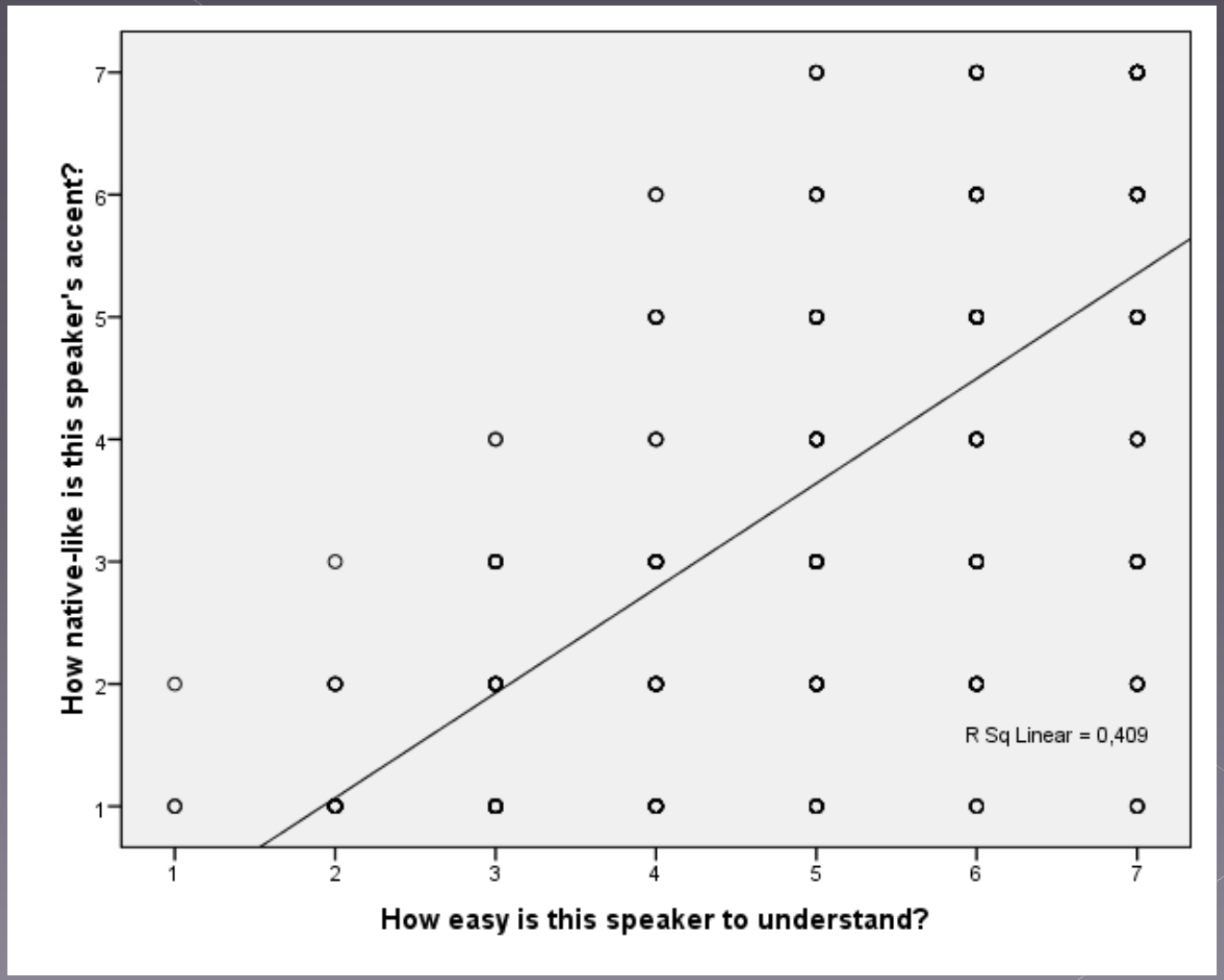
- Is there a correlation between how native-like and how comprehensible a speaker is?

Correlations

			How native-like is this speaker's accent?	How easy is this speaker to understand?
Spearman's rho	How native-like is this speaker's accent?	Correlation Coefficient	1,000	,638**
		Sig. (2-tailed)	.	,000
		N	464	463
	How easy is this speaker to understand?	Correlation Coefficient	,638**	1,000
		Sig. (2-tailed)	,000	.
		N	463	463

**. Correlation is significant at the 0.01 level (2-tailed).

Graph Correlation 3



Correlation 4

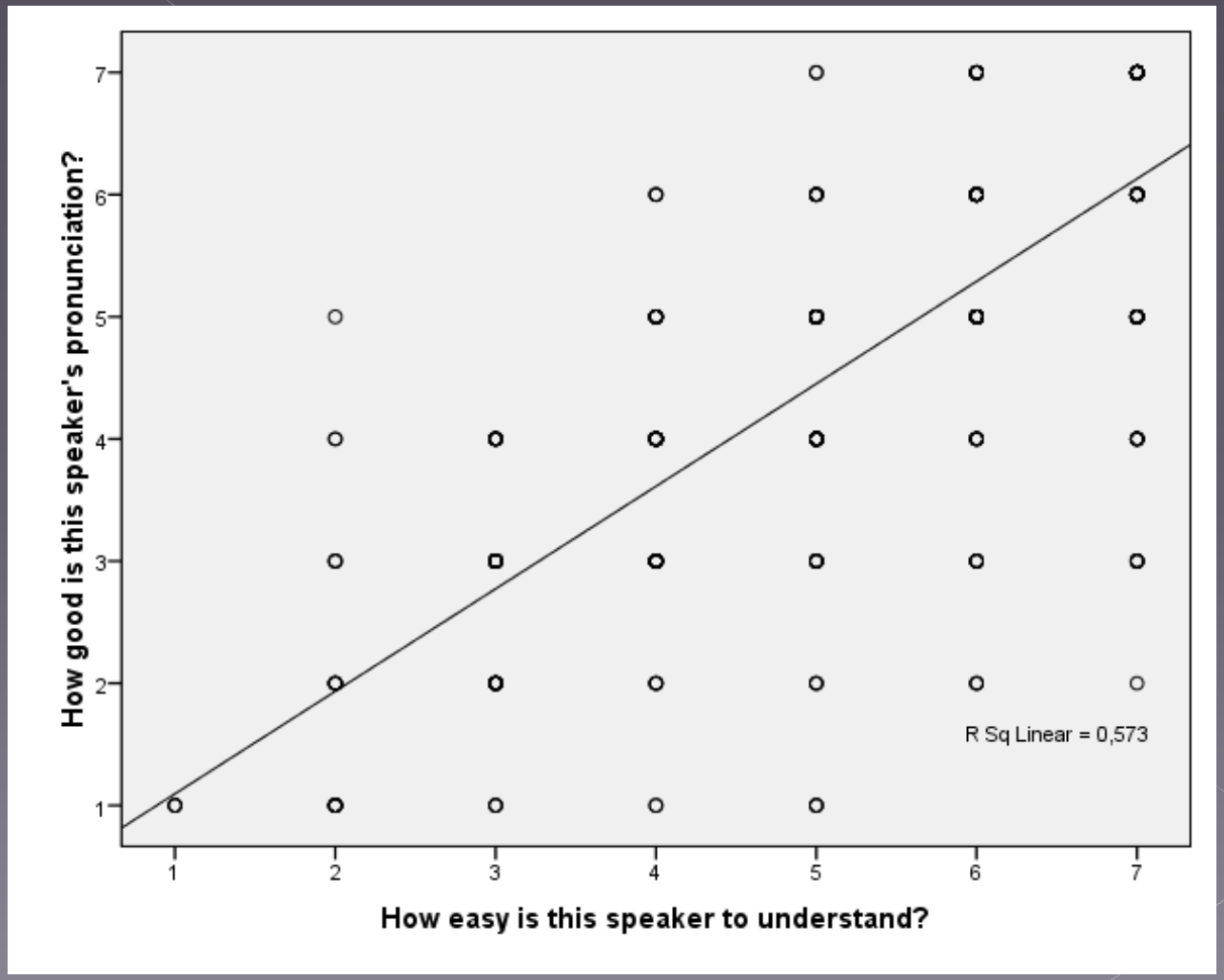
- Is there a correlation between comprensibilty and pronunciation?

Correlations

			How easy is this speaker to understand?	How good is this speaker's pronunciation?
Spearman's rho	How easy is this speaker to understand?	Correlation Coefficient	1,000	,752**
		Sig. (2-tailed)	.	,000
		N	463	463
	How good is this speaker's pronunciation?	Correlation Coefficient	,752**	1,000
		Sig. (2-tailed)	,000	.
		N	463	463

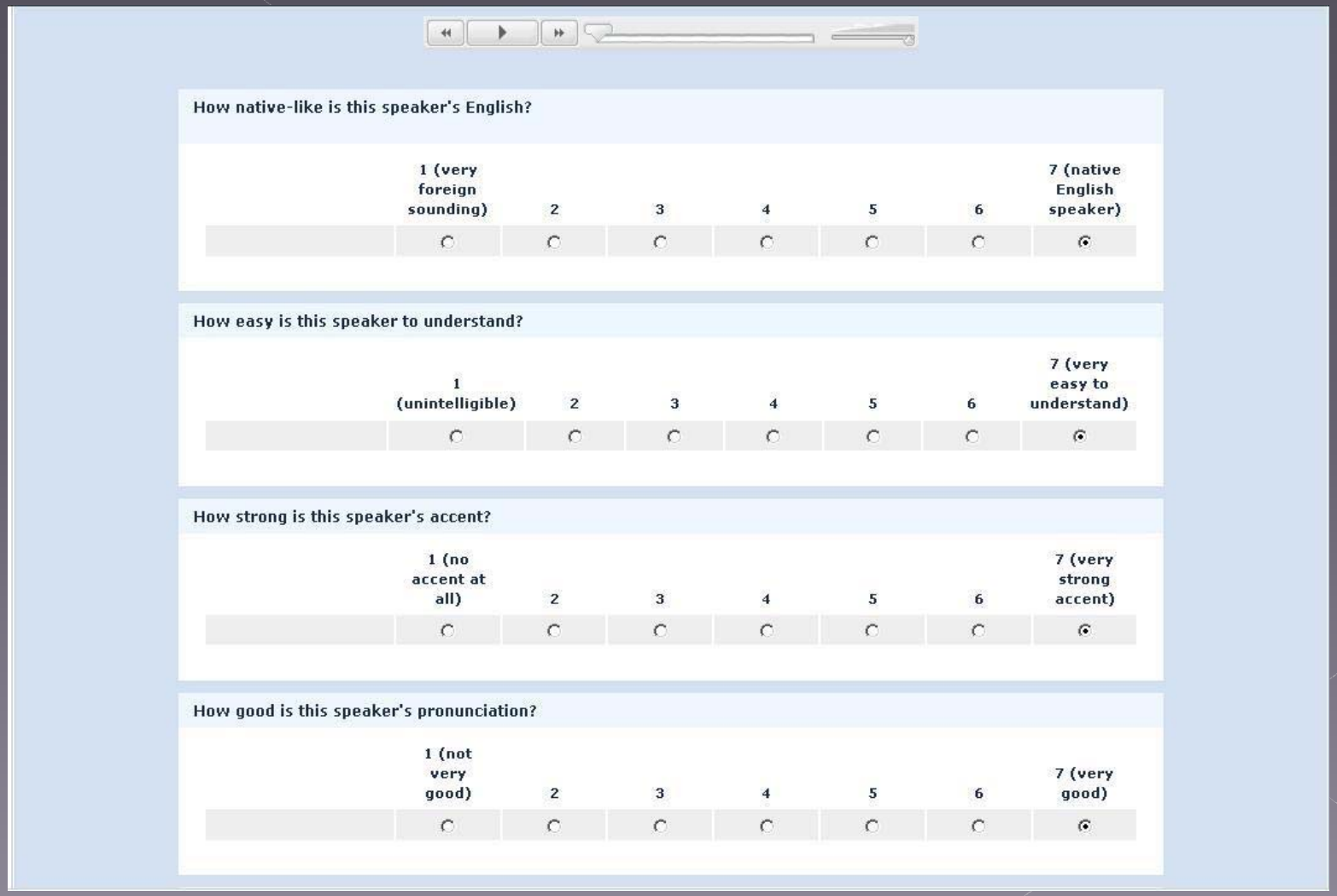
**. Correlation is significant at the 0.01 level (2-tailed).

Graph Correlation 4



Correlations

- Need to clean up the data



How native-like is this speaker's English?

	1 (very foreign sounding)	2	3	4	5	6	7 (native English speaker)
	○	○	○	○	○	○	◉

How easy is this speaker to understand?

	1 (unintelligible)	2	3	4	5	6	7 (very easy to understand)
	○	○	○	○	○	○	◉

How strong is this speaker's accent?

	1 (no accent at all)	2	3	4	5	6	7 (very strong accent)
	○	○	○	○	○	○	◉

How good is this speaker's pronunciation?

	1 (not very good)	2	3	4	5	6	7 (very good)
	○	○	○	○	○	○	◉

Questionnaire Design

- Selection of samples
  - Why include native-speakers?
  - What's most important to control for?
- Why the memory task?
- Which questions to use?

Gathering Data

- 100-200 participants
- Incomplete questionnaires
  - More data for the samples that appear on the first half than on the second half

Additional Research Questions

- Which non-native speakers are most native-like?
- Are people better at identifying their own accent and accents that they are more familiar with (due to where they live)?
- How good are people in general at identifying where a speaker is from and what his/her native language is?
- Which accents are generally easiest to identify?