# PHONOLOGICAL (DIS)SIMILARITY

## REDUPLICATION, CONFUSABILITY, AND THE LEXICON IN BENGALI

SAMEER UD DOWLA KHAN, REED COLLEGE

GLOSS COLLOQUIUM, 5 FEBRUARY 2016

# OUTLINE

- Identity and similarity in phonology

- Identity avoidance, with a puzzle from English
- Production data from Bengali
- Gradient similarity avoidance

- Shared natural classes
- Weighted shared natural classes

- Lexical statistics
- Perceptual confusability

Many processes incorporate

- 
- in Chumash (Beeler 1970)
  - [ki kín] + [us] " *[ki kínus] " [kiskínus] ‘I saved it for him’
- (identity avoidance)
  - merry + -ly " merrily
  - silly + -ly " *sillily

But some phenomena in perception and the lexicon are best described as involving

- Lexical effects in Muna (Coetzee & Pater 2005)
  - [d] is found in fewer roots with [t] than with [n]
- Perceptual

- [d kt $m_F$ kt ] ‘doctor$_{\text{DISMISSIVE}}$’
- As opposed to [d kt d kt ] ‘(real/prototypical) doctor’
- Most common in lgs across southern Asia

## Phonological properties

- Total reduplication
-

Unlike prototypical reduplication, echo reduplication typically the base and RED to be

- Unlike “emergence of the unmarked” cases of base-RED nonidentity, e.g. Sanskrit (Steriade 1988)
-

Through various means, lgs work to
between base and RED in echo forms

Survey of echo forms in >100 lgs of India found
in every case (Trivedi 1990)

Previous workk21 Tf [ (08.0401cm7eoWrq 1 0 0 1i) -48 (!) Tj i!cs 0.32

What about English [ $m_F$]?

- [d kt ] ‘doctor’ " [d kt $m_F$ kt ] ‘doctor$_{DISMISSIVE}$’
- [skul] ‘school’ " [skul $m_F$ul] ‘school$_{DISMISSIVE}$’

Online survey, 190 respondents (Nevins & Vaux 2003)

: 95–97% of speakers rejected echo forms with [ $m_F$] for the 3 [ m]-initial words

- [ muz] ‘schmooze’ " *[ muz $m_F$uz] ‘schmooze$_{DISMISSIVE}$’

Interestingly, 30% of speakers also rejected echo forms with [ m

Possible explanations:

The “two dialects” possibility

- 65% of subjects obey
- 30% obey , where [ n] and [ m] are of the same category: “sounds similar to [ $m_F$]”

The “matter of degree” possibility

- 95% obey , of whom:
- 65% considered [ n] and [ $m_F$] are sufficiently dissimilar

# A PUZZLE FROM ENGLISH

Another possible explanation: “this isn’t English”

- and possibly to the language
- in English than in other lgs
- [ m] is , restricted to from Yiddish
- Construction is possibly borrowed from Yiddish (Southern 2005)

# MOTIVATION

To understand if echo reduplication can employ gradient similarity avoidance, we need a lg in which:

- Echo reduplication is a fully , feature
- The fixed segment is a relatively
- The fixed segment has many sounds

- Default fixed segment [$t_F$][2]: crosslinguistically unmarked
- [t] has high token freq. (definite marker & classifier [-ta])
- Attested backup fixed segments [$m_F$ $f_F$ $p_F$ $u_F$] (Ray et al. 1966)
- Inventory has many [t]-like sounds: [t d d t t d t s...] (Khan 2010)

[1] Specifically, urban colloquial Bangladeshi varieties
[2] [t t d d ] can be retroflex in Bengali, but are typically alveolar in these varieties (Khan 2010)

# QUESTIONS

Does echo reduplication in Bengali involve…

- ,
- , or
- ?

If it is the latter, how can similarity be objectively on a gradient scale?

As a comparison, we can investigate other parts of Bengali phonology that expected to employ this gradient similarity:

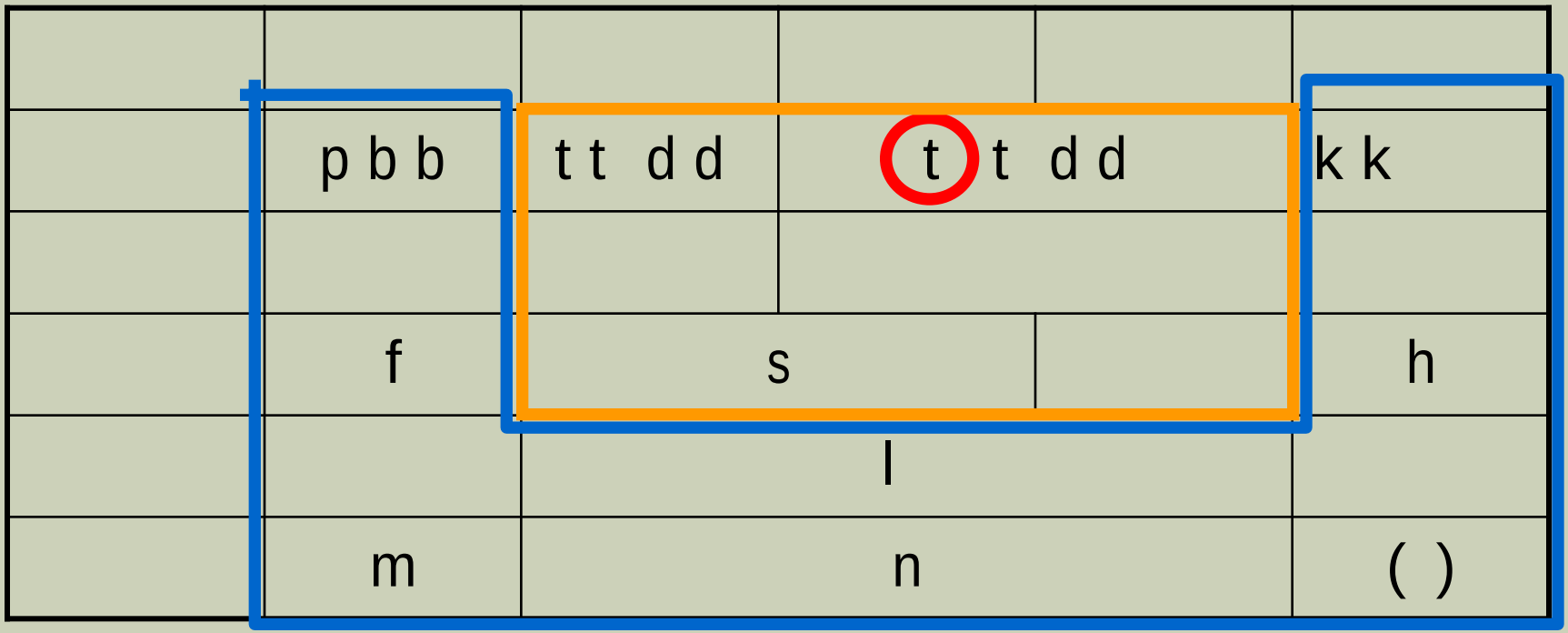
- Lexical restrictions
- Perceptual

## words

- Disyllabic stems
-

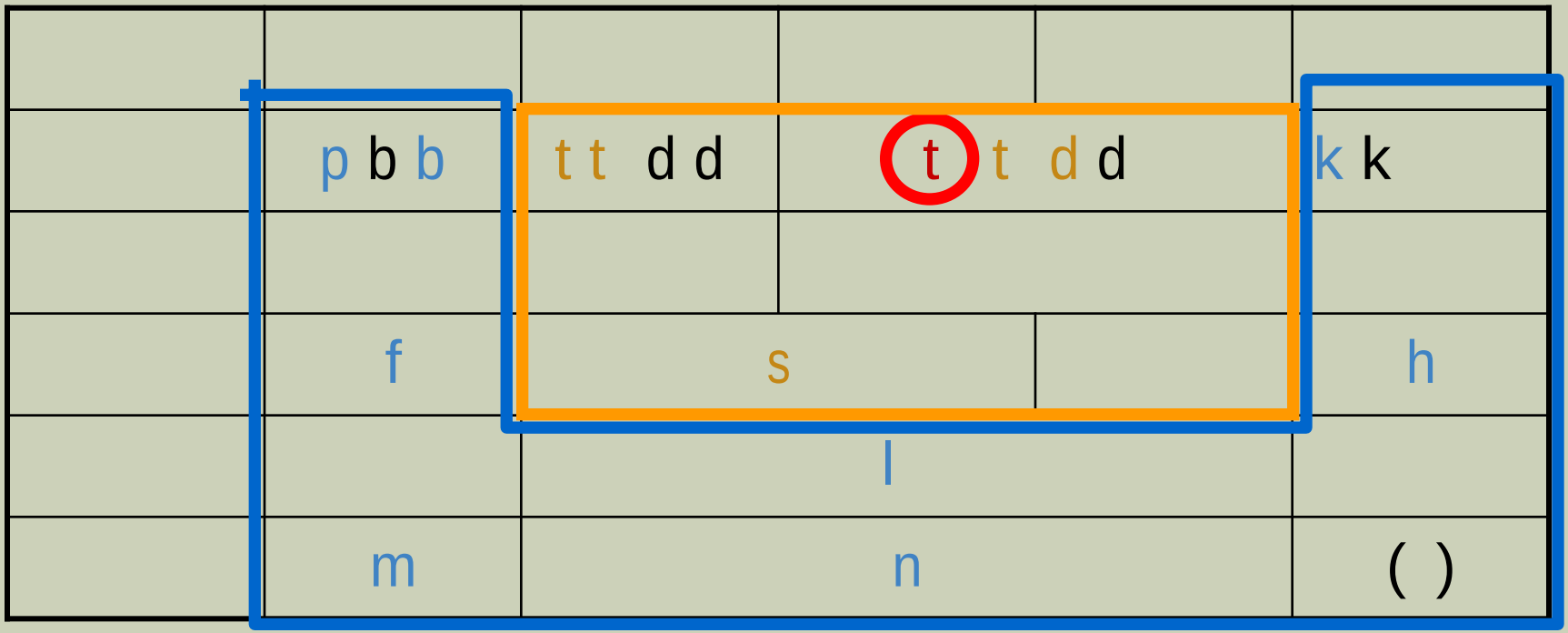
# EXPERIMENT I: STIMULI

Consonants of Bangladeshi Standard Bengali (Khan 2010)



# EXPERIMENT I: STIMULI

Consonants of Bangladeshi Standard Bengali (Khan 2010)



of Bengali

- Varied dialect background
- Residents of CA
- Paid $10

-

words will never use [$t_F$]

words will always use [$t_F$]

words are what are being tested:

- Hypothesis 1: = (categorical identity)
- Hypothesis 2:

words will never use [$t_F$]

words will always use [$t_F$]

words are what are being tested:

- Hypothesis 1: = (categorical identity)
- Hypothesis 2: = (categorical similarity)
- Hypothesis 3: is on a continuum

*[t...$t_F$] = *[t ...$t_F$] [

words will never use [$t_F$]

words will always use [$t_F$]

# EXPERIMENT I: RESULTS

was borne out

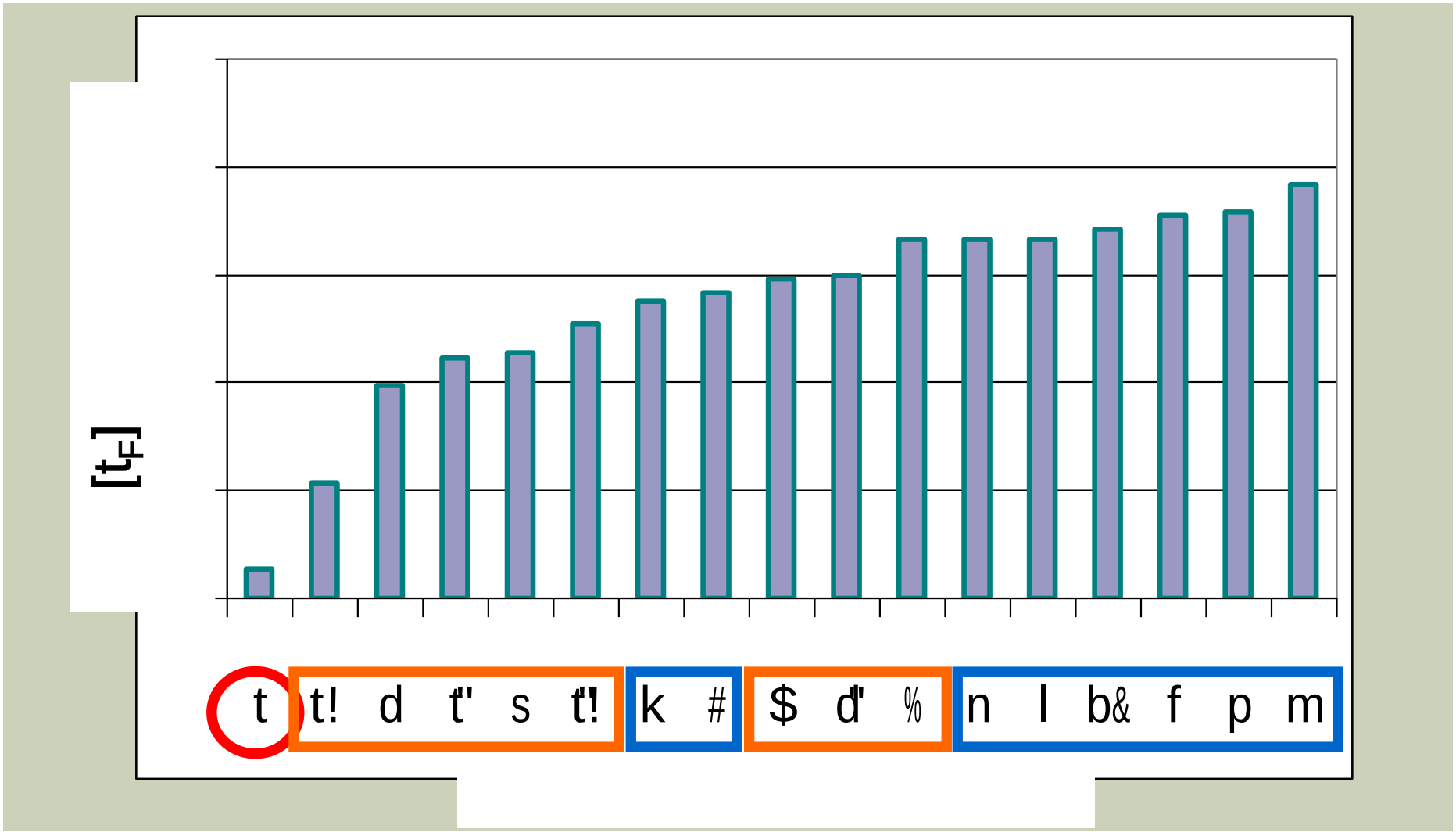
Similarity words lie on a continuum

- Disprefer [$t_F$] but not outright ungrammatical
- Some consonants are more [t]-like in behavior than others

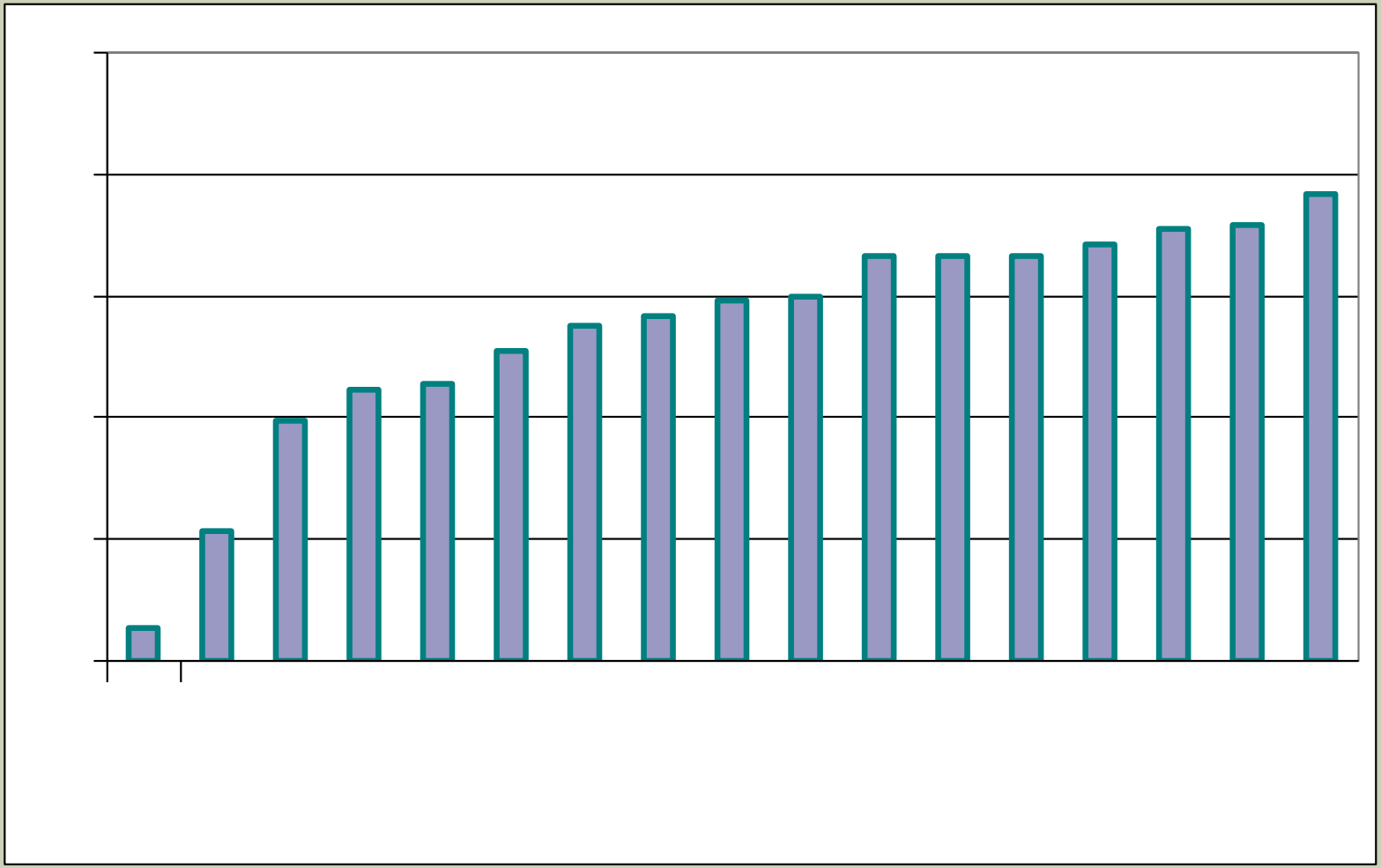
Seems like [$t_F$]

[t]

# EXPERIMENT I: RESULTS



# EXPERIMENT I: RESULTS



# EXPERIMENT I: DISCUSSION

Echo reduplication in Bengali appears to incorporate a notion of

- No straightforward clustering of consonants
- Heavy overlap across clusters
- Like the “matter of degree” hypothesis from English puzzle

We should confirm our suspicion that our reduplication data can be modeled on an

Is there a that Bengali speakers are using to calculate the similarity of an initial C and [t]?

# SNC: METRIC

In the SNC metric, similarity of $C_1$ and [t] is quantified as:

$$sim(C_1, t) = \frac{\#\text{ natural classes shared by } (C_1, t)}{\#\text{ shared natural classes} + \#\text{ non-shared natural classes}}$$

Compared SNC-similarity (line) to Exp 1 results (bars)

The SNC metric does an okay job overall ($r^2$ = .584)

However, the area where it crucially to predict the data is the (coronal obstruents)

The metric treats [t] as inherently more similar to [t]

Original SNC metric derives directly from the phoneme inventory and feature set

But what if we maintain the basic model but incorporate
?

Let’s try a little

Weighting [dist] our4 ] itur4 -42( ) -36 ([) -36 (]) -36 (n) -36 ( ) -36 (t

# WEIGHTED SNC: METRIC

In an SNC-like model with feature , similarity of $C_1$ and [t] is quantified as follows: (Wilson, p.c.)

$$sim(C_1, t) = \exp\left(-\sum_{i=1}^{\#features} \mathbf{i}(1 - {}_i(C_1, t))\right)$$

= weight of the feature $f_i$

${}_i(C_1, t)$ = 1 (feature value shared) or 0 (not shared)

Where weights are drawn from the variation in the reduplication results, as follows:

Probability of $[t_F]$ use in the RED of a base with initial $C_1$

$$P = ((m!) \div (n!(m\ \ n)!)\ (1\ \ sim(C_1, t))^n\ (sim(C_1, t))^{m\ \ n}$$

P = probability that $C_1$-initial base will be reduplicated with $[t_F]$ *n* times out of a total of *m* trials

*m* = number of reduplications for C

# WEIGHTED SNC: DISCUSSION

With , the SNC metric can closely model the reduplicative data ($r^2 = .855$)

- [voice]: .554
- [distributed]: .400
- [strident]: .249
- [spread glottis]: .198
- All other features have a weight of 0.100

# NEW QUESTION

Okay, but have we compromised the model?

Is it no longer a similarity metric, but just a model of the reduplicative data?

Let's see if our reduplicative data resemble other areas where gradient, lg-specific similarity is arguably relevant:

- Lexical (McCarthy 1994)
- Perceptual (Shepard 1972)

# COOCCURRENCE

of two Cs is often with their within roots (Greenberg 1950)

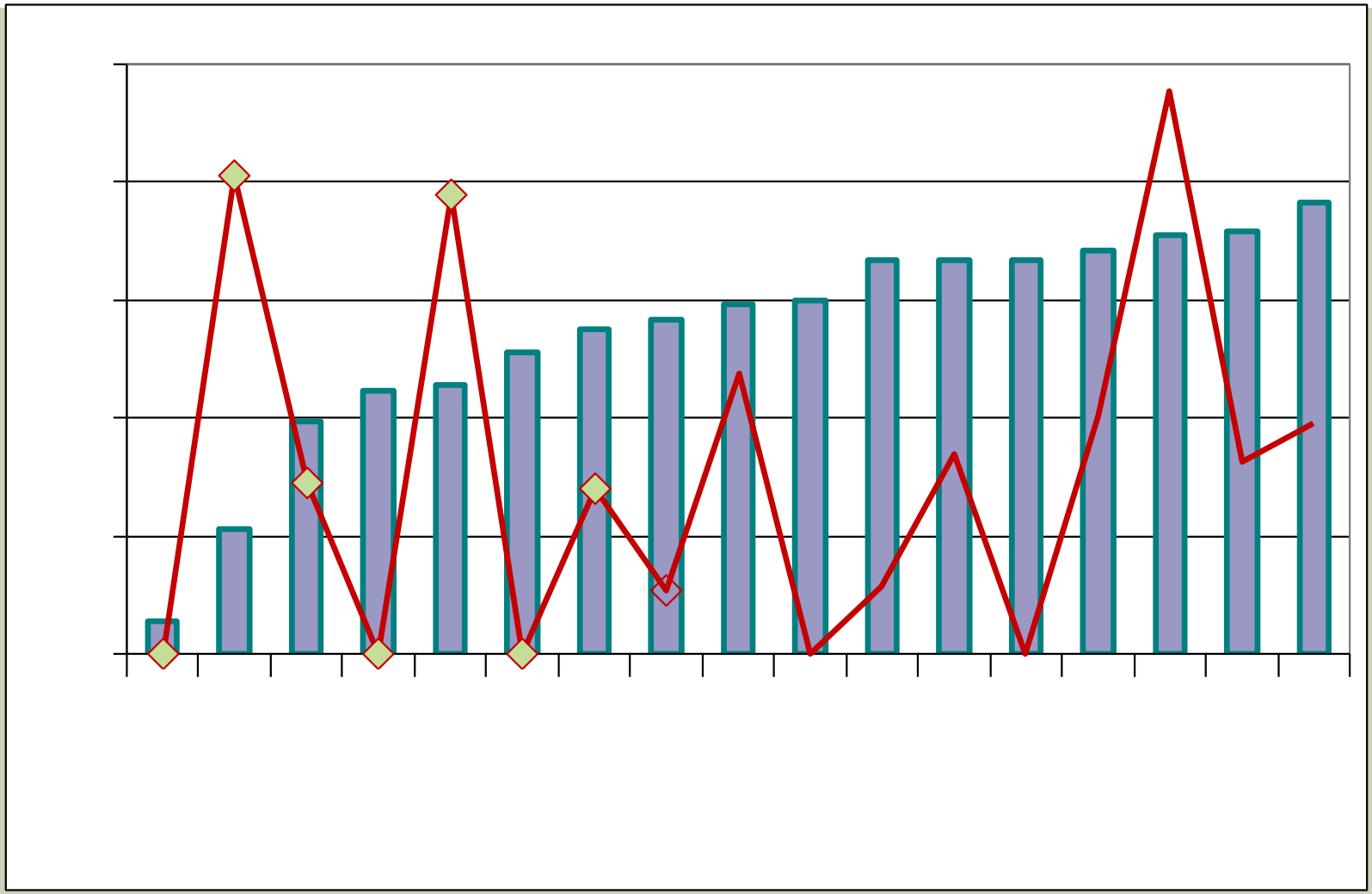
- English: two LAB or two DOR are underattested in [sCVC]: skip, speak, skim, smack..., *smap, *scog, *spobe, *speam (Fudge 1969)
- Arabic: velars & uvulars rarely cooccur within roots (Frisch et al. 2004)

: the less often a C cooccurs with [t] in a root, the less often it will take [$t_F$] in its echo RED

If we see a strong correlation with the reduplicative data, this could be independent support for our weighted model

Similarity of initial $C_1$ and medial [t] is the inverse of their observed / expected lexical cooccurrence: (Frisch et al. 2004)

Examined the cooccurrence of all initial Cs with medial [t]



# COOCCURRENCE: DISCUSSION

The lexical cooccurrence model of similarity
the observed [$t_F$]-avoidance patterns ($r^2 = .004$)

Possible explanations:

Lexical cooccurrence in Bengali involves similarity, but echo reduplication does not (unlikely, see results)

Lexical cooccurrence in Bengali does not involve similarity, while echo reduplication does (possible)

Corpus had 865 CVCV roots; 64 with medial [t]

! cf. Arabic corpus of 2674 roots (Frisch et al. 2004)

The other area to look for the effects of gradient similarity is in

- Hindi: [ ] is misidentified as

# EXPERIMENT II: SETUP

(MFC) listening experiment

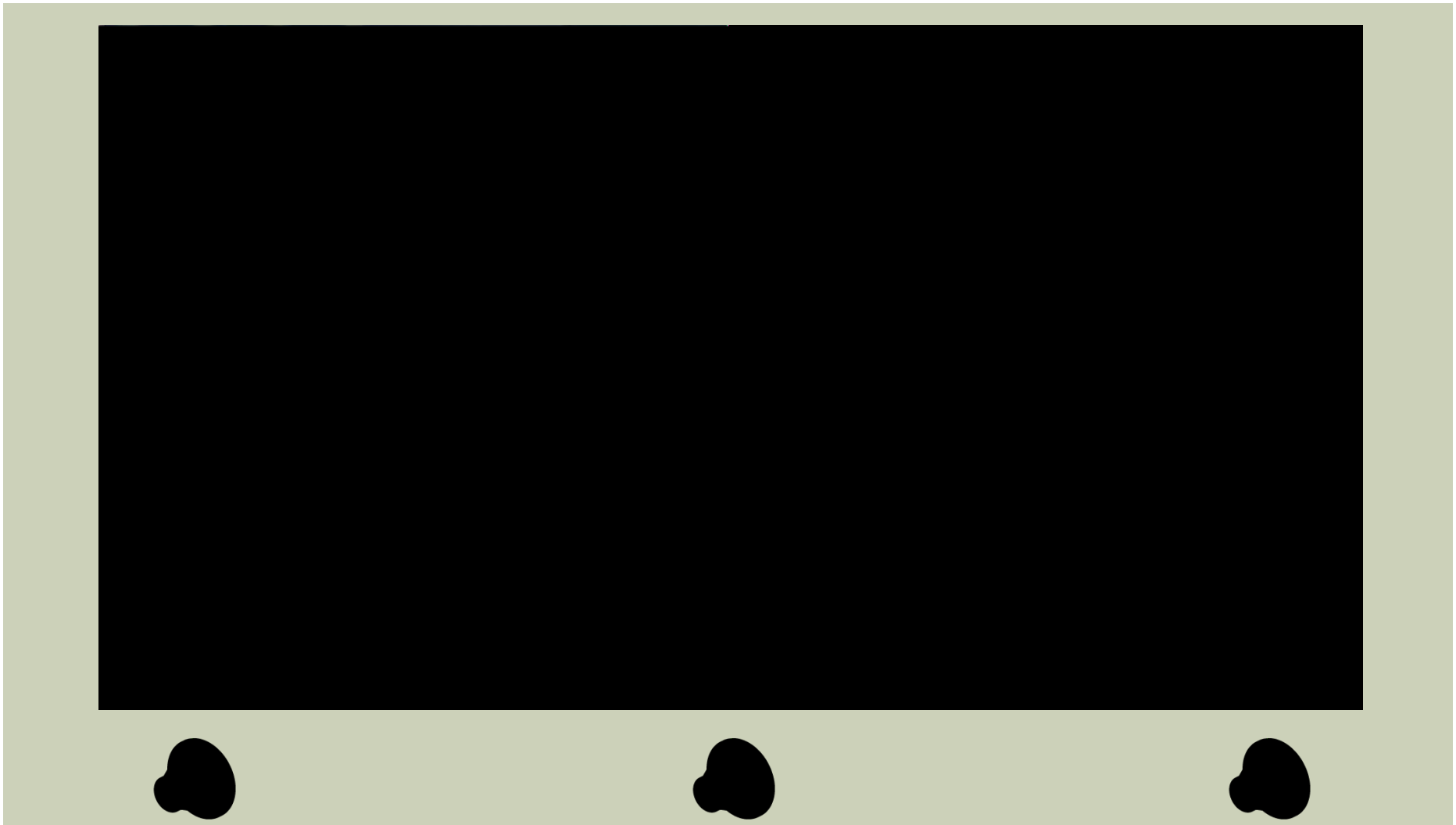
- Participants identify the consonant they hear
- Run in Praat (Boersma & Weenink 2013)
- Sony MDR-V200 headphones connected to laptop
- Experiments took place in quiet room in participants' homes

of Bengali (13F, 12M)

- Reported no hearing difficulties
- Varied dialect background
- Residents of or visitors to CA
- Paid $20

- Onsets: 27 legal [Ca] syllables (all Cs but [ ])
-

# EXPERIMENT II: TASK



The C most confused with [t] should be [t ]

- Generalized: should be the feature

Next most confused with [t] should be [d]

- Generalized: should be the feature

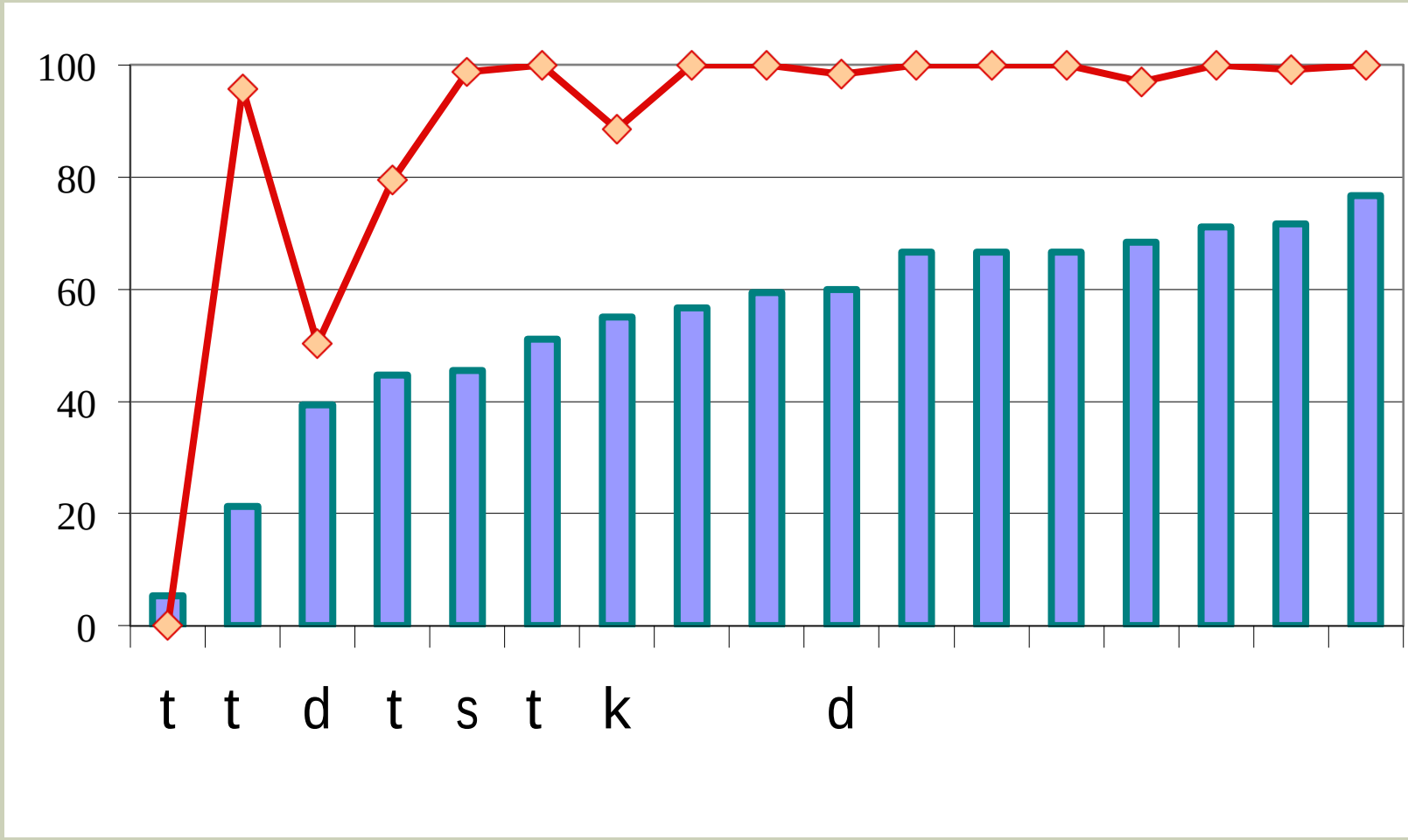
# CONFUSABILITY: METRIC

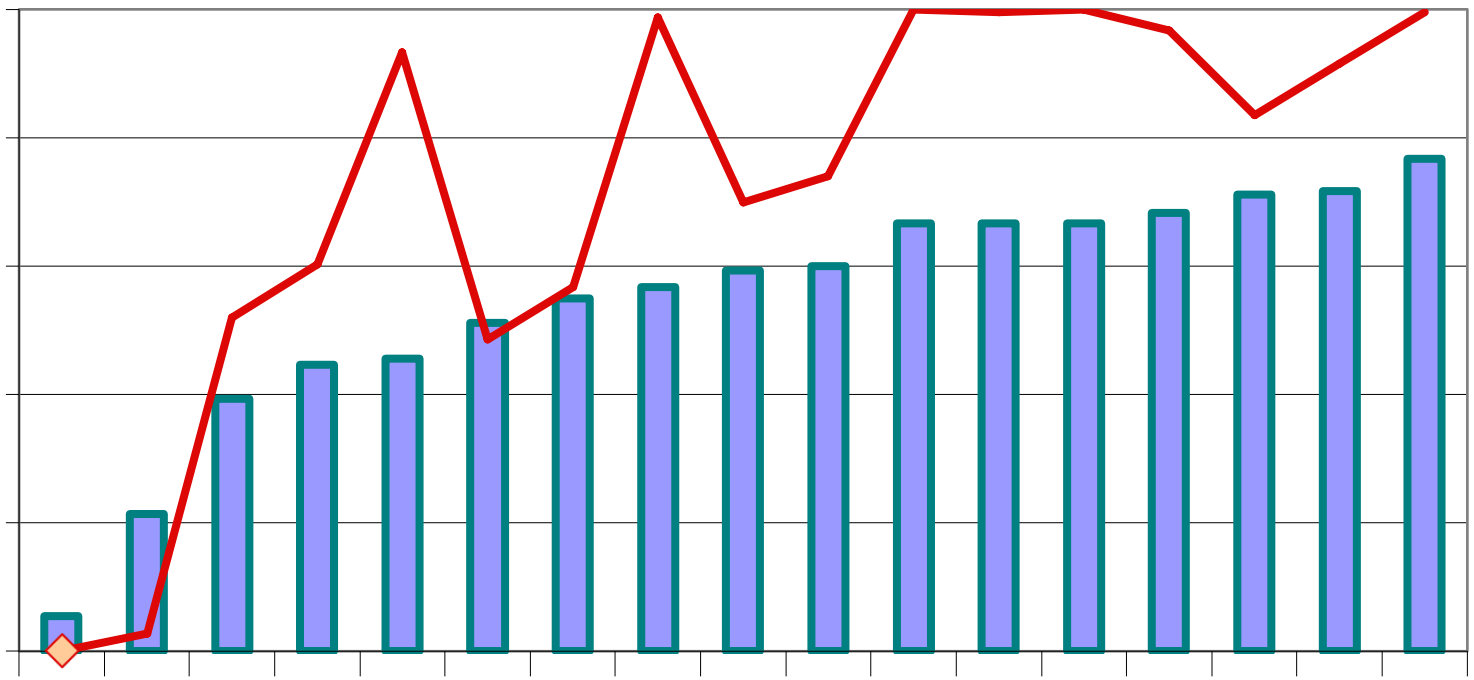
Similarity of $C_1$ and [t] as drawn from confusion rate is quantified as follows: (Shepard 1972)

$$sim(C_1, t) = \frac{\#(C_1{:}t) + \#(t{:}C_1)}{\#(C_1{:}C_1) + \#(t{:}t)}$$

Compared Exp 2 perceptions to Exp 1 productions

- Removed “quiet” condition results (at ceiling)
- Looked at onsets and codas separately


100
80
60
40
20
0
t t d t s t k d



# SYNTHESIS OF RESULTS

Okay, we need a recap.

What did we do again?

- Task 1: examine in echo reduplication
- Task 2: establish that fixed segment choice is
- Task 3: improve the SNC in a thought experiment with
- Task 4: find no correlation with
- Task 5: find significant correlation with

# CONCLUSIONS

The current study demonstrates that fixed segment choice in Bengali echo reduplication is

I argue that the choice of fixed segment involves a systematic avoidance of , because:

- The patterns are (partially) predicted by the
- The patterns correlate with (in codas)

The patterns clearly show that this similarity is

Echo reduplication is one of many phenomena previously treated as categorical but more recently seen as gradient

!

# CONCLUSIONS

The current study proposes a modified version of the SNC metric of similarity

for lg-specific application in diverse phonological phenomena

The study also provides an interesting case in which the SNC metric can measure similarity in phonological phenomena *other than* lexical cooccurrence effects

# REMAINING QUESTIONS

Is Bengali echo reduplication a special case, or should we look for gradient similarity in many more lgs?

Why are the lexical cooccurrence effects of Bengali so different from the reduplicative results?

How does this change as speakers deal with multiple phoneme inventories, e.g. bilinguals?

Significant features in the perception of (Hindi) consonants. *JASA* 45(3).

Reduplication with fixed segmentism. *Linguistic Inquiry* 30(3), 327-364.

Sibilant harmony in Chumash. *IJAL* 36(1), 14-17.

*The Classical Tibetan language.* SUNY Press.

Praat: doing phonetics by computer, version 5.3.39, http://www.praat.org/

Gradient phonotactics in Muna and Optimality Theory. Ms.

Patterns of English phoneme confusions by native and non-native listeners. *JASA* 116(6), 3668-3678.

Similarity avoidance and the OCP. *NLLT* 22, 179-228.

Syllables. *Journal of Linguistics* 5, 253-287.

The patterning of root morphemes in Semitic. *Word* 6, 162-181.

The phonetics and phonology of Semitic pharyngeals. In Keating, Patricia (ed.) *Phonological structure and phonetic form: papers in laboratory phonology 3*, 191-233.

Logical and substantive scales in phonology. UC Berkeley PhD dissertation.

Metalinguistic, shmetalinguistic: the phonology of shm-reduplication. Ms.

*Bengali language handbook.* Washington: Center for Applied Linguistics.

Psychological representation of speech sounds. In David, E. E.; Denes, P. B. (eds.) *Human communication: a unified view*, 67-113.

*Contagious couplings: transmission of expressives in Yiddish echo phrases.* Westport: Greenwood.

Reduplication and syllable transfer in Sanskrit and elsewhere. *Phonology* 5(1), 73-155.

Vowel harmony in Turkish and Turkmen. Ms.