13th Meeting of the ALT - Università degli Studi di Pavia – September 4-6 2019

An empirical reconsideration of tone volatility and segment stability in Mixtec languages

SANDRA AUDERSET[1,2], CHRISTIAN D. BRENDEL[1],
ERIC W. CAMPBELL[1], SIMON L. PETERS[1],
KEVIN SCHAEFER[1] & ALBERT VENTAYOL-BOADA[1]

1 UNIVERSITY OF CALIFORNIA, SANTA BARBARA
2 MPI FOR THE SCIENCE OF HUMAN HISTORY, JENA

Introduction

Tones are often claimed to be volatile:

- *"The complex composition of a tone, as much as its phonological environment, gives ample opportunity for tones to change (...)"* (Ratliff 2015:254)
- *"Somewhat unusually typologically, even closely related dialects in PNG may have quite different prosodic systems."* (Cahill 2011:19)

In another sense, tone--or rather, *tonality*--is often regarded as a stable typological feature (cf. Nichols 1995; Wichmann & Holman 2009; among others)

Defining tone instability

Statements about tone (in)stability can be interpreted in terms of:

- as already mentioned, tone as a typological feature (presence vs. absence of tone)
- changes in tone (or tone melody) inventories
- tonal phonology (distribution, processes)
- range of phonetic realizations of a toneme

Tone change remains poorly understood

- tone is under-studied generally, but especially in historical linguistics (Janda & Joseph 2003)
- focus on tonogenesis rather than tone change (Haudricourt 1954; Hombert et al. 1979; Kingston 2011; Thurgood 2002)
- tone seen as unruly/difficult and often excluded (Welmers 1959)
- the lack of studies on tone change hampers our understanding of the prehistory of tone languages (Campbell in press)
- over half of the world's languages are tonal (Yip 2002)

Motivations for this study

- empirically-based test of the notion that tone is volatile, or more prone to change, relative to segments
- this raises issues of comparability between tone and segments
- present the first results of an ongoing study of tonal and segmental change in Mixtec languages (Otomanguean. Mexico), spoken in the Ventura County, CA diaspora
- tone has also been reported to be markedly volatile in Mixtec:

  "*tone is among the first features to vary between towns speaking similar varieties of Mixtec*." (Josserand 1983: 243)

Thanks

We would like to thank our collaborators in Ventura County, CA, for their continued support and for the permission to share their data here today:

- Griselda Reyes Basurto
- Carmen Hernández Martínez
- Gabriel Mendoza
- Juvenal Solano
- Yésica Ramirez

Roadmap

1. Brief introduction to Mixtec
2. Methods
3. Results from pilot study
4. Summary and further research

A brief introduction to Mixtec

- Otomanguean > Eastern Otomanguean > Amuzgo-Mixtecan > Mixtecan
- Mixtec language group
  - 52 languages (Egland 1983; Lewis et al. 2016)
  - 81 varieties (INALI 2009)
  - 12 primary subgroups (Josserand 1983)
- complex systems of grammatical and lexical tone
- tone-bearing unit is the mora

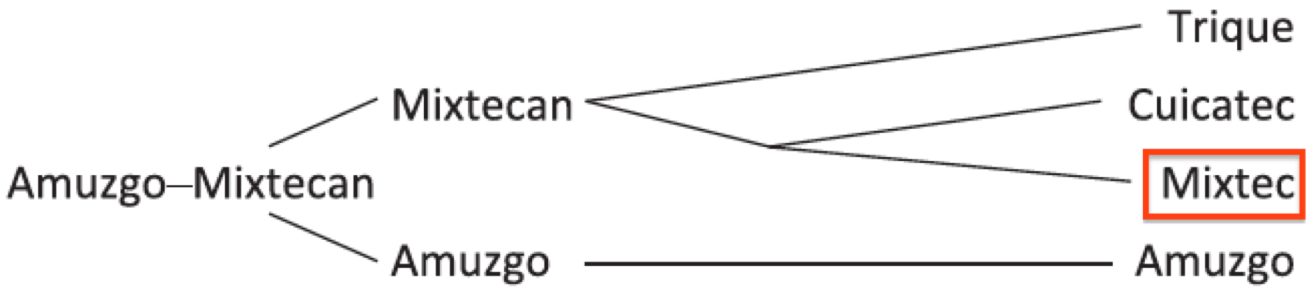
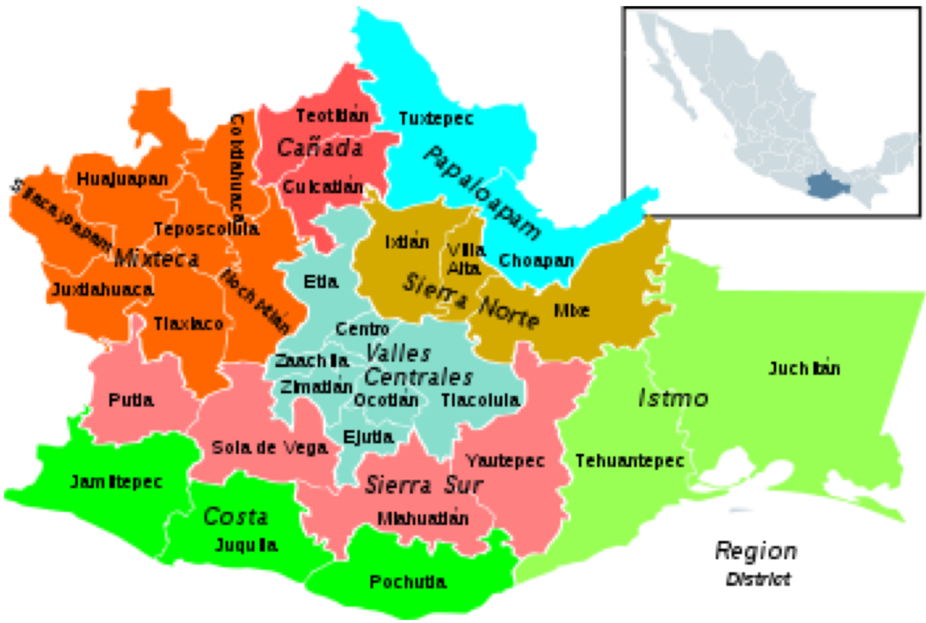


Image: Amuzgo-Mixtecan branch of the Otomanguean stock (adapted from Campbell 2017:3)

The speech communities



Southern Mexico

Ventura County, California, USA

Geographical location of the 5 varieties

Data collection

- Manual data collection
  - community-based documentation (Cruz & Woodbury 2014b, Czaykowska-Higgins 2009) in Oxnard, CA through partnership with the Mixteco/Indígena Community Organizing Project (MICOP)
  - multi-variety lexical database with entries in the practical orthography developed with community members
- Computer-generated data
  - regularization of the orthography
  - orthography to IPA conversion
  - tone melody extraction

Study design

- Dataset:
  - 330 cognate sets from 5 varieties
  - Aligned with a modified version of LingPy (List et al. 2018)
  - Tones are compared (not tonal melodies)

Team member	Location in MX: Village (State)	Dialect area (Josserand 1983)	Endonym
Griselda Reyes Basurto	Tlahuapa (Guerrero)	Guerrero	Tù'un Sàví
Gabriel Mendoza	Piedra Azul (Oaxaca)	Southern Baja	Tù'un Ntá'vi
Carmen Hernández Martínez	San Martín Duraznos (Oaxaca)	Southern Baja	Tù'un Ntá'ví
Juvenal Solano	San Sebastián del Monte (Oaxaca)	Central Baja	Tù'ùn Nta'ví
Yésica Ramírez	San Juan Mixtepec (Oaxaca)	Mixtepec	Tù'ùn Ndá'vi

Study design

- Metrics computed:
  - Entropy
  - Simple Levenshtein distance
  - Feature-based Levenshtein distance
- Analyses:
  - Multinomial models (Tone vs. Vowels vs. Consonants)
  - Logistic regression (Tone vs. Segments)

Example cognate sets

'wing'

```
CogID #24
n   d   i   `   ʃ   ĩ   ¯
n   ʤ   i   `   ʃ   ĩ   ˇ
n   d   i   `   s   i   `
n   ʥ   i   `   ɕ   ĩ   `
n   ʣ   i   `   ʦ   i   ˇ
```

'smoke'

```
CogID #63
j   u   `   ʔ   m   a   `
–   i   ¯   ʔ   m   a   `
ɲ   i   `   –   m   a   `
ɲ   u   `   ʔ   m   a   `
ɲ   u   ¯   ʔ   m   a   ¯
```

Entropy

- Entropy = a measure of the 'disorder' in a system
- Calculated for aligned correspondence sets
  - Entropy algorithm from Chao & Shen (2003)
- Assumption: entropy = 0 at the point of divergence (proto-language)
- Entropy could take phonetic or phonological representation as input (but phonetic input would inflate differences in predictable phonology)

CogID #24

n	d	i	ˋ	ʃ	ĩ	ˉ
n	ʤ	i	ˋ	ʃ	ĩ	ˇ
n	d	i	ˋ	s	i	ˋ
n	ʥ	i	ˋ	ɕ	ĩ	ˋ
n	ʣ	i	ˋ	ʦ	i	ˇ
0	2.28	0	0	2.28	0.70	1.36

Distance

- Levenshtein distance = the minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into another

CogID #63

j	u	`	ʔ	m	a	`	
–	i	¯	ʔ	m	a	`	
1	1	1	0	0	0	0	**= 3**

- Levenshtein distance approximates number of steps in diachronic divergence, especially if multiple features are considered

Simple vs. Feature-based units

- simple units: each segment/toneme is considered as one unit
- feature-based units: each segment is considered as made up of multiple features
- example:

	Simple		Feature-based (simplified matrix)	
	/d/	/d/	[+voice,+alveolar] (/d/)	[+voice,+alveolar] (/d/)
	/k/	/g/	[-voice,+velar] (/k/)	[+voice,+velar] (/g/)
Distance	**1**	**1**	**2**	**1**

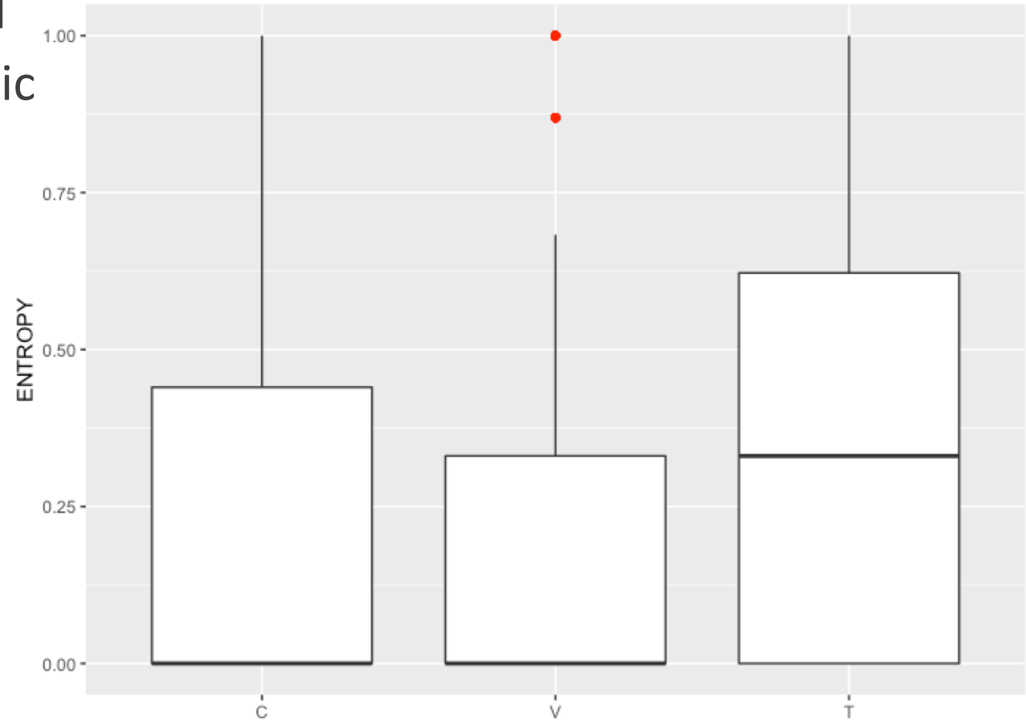
Scaling options

Recap: tone is not directly comparable to segments

- estimate the range of variability for segments vs. tones and resize metrics
- compare features to each other and treat tone as a single feature
- compare tone only to something with comparable variability (like vowel height, with three distinctive levels)
- ignore the multidimensional nature of segmental change (classic Levenshtein)

Differences in simple entropy

- Calculation:
  - Chao-Shen algorithm
  - entropy values normalized by maximum possible entropic state for system

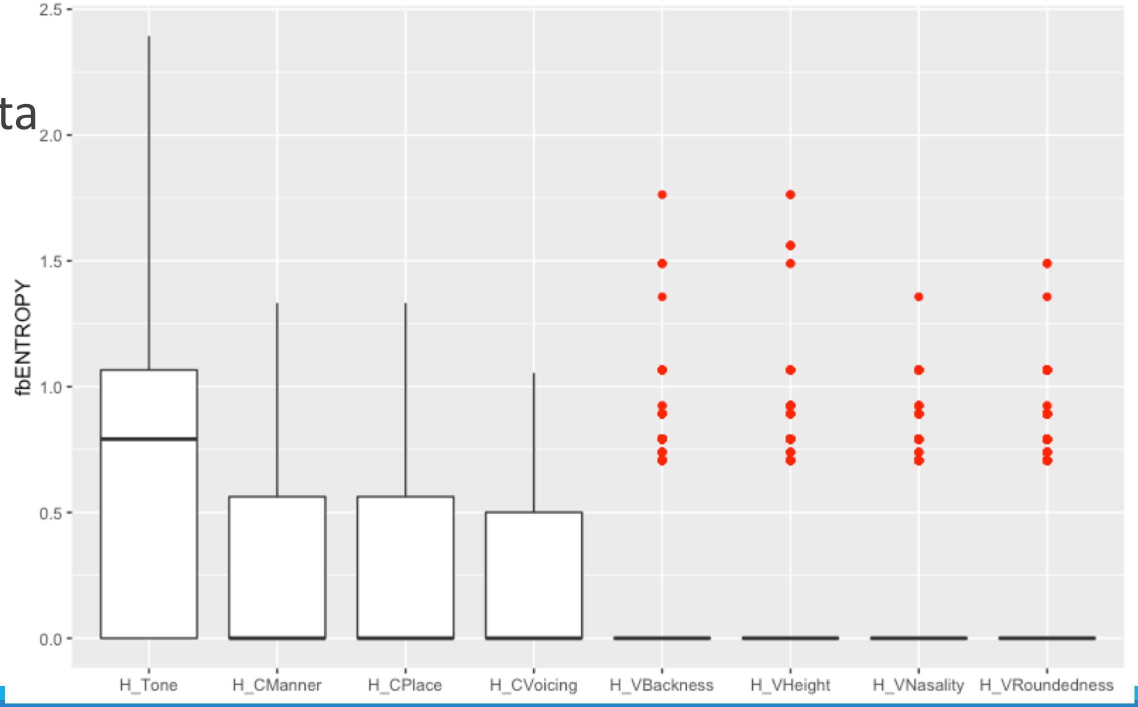


Differences in simple entropy

- Multinomial regression model:
  - formula: {tone | vowel | consonant} ~ entropy
  - result: highly significant
  - interpretation: in our data set, tone shows more entropy than both vowels or consonants
- Logistic regression model
  - formula: {tone | segment} ~ entropy
  - result: highly significant
  - interpretation: tone also shows more entropy vs. segments as a whole

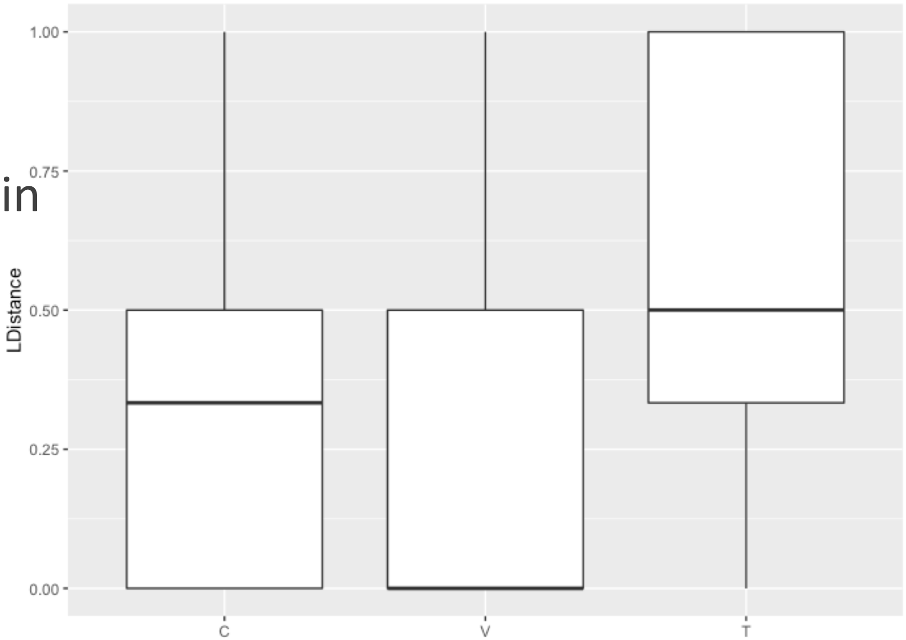
Differences in feature-based entropy

- Calculation: Chao-Shen algorithm
- Multinomial regression model:
  - formula: {each feature} ~ entropy
  - result: highly significant
  - interpretation: in our data set, tone shows more entropy than each feature of consonants and vowels



Differences in simple Levenshtein distance

- Calculation:
  - penalty for addition, deletion, change: 1 edit/word
- Multinomial regression model:
  - formula: {tone|vowel|consonant} ~ S.Levenshtein
  - result: highly significant
  - interpretation: in our data set, tone shows more distance than both vowels or consonants

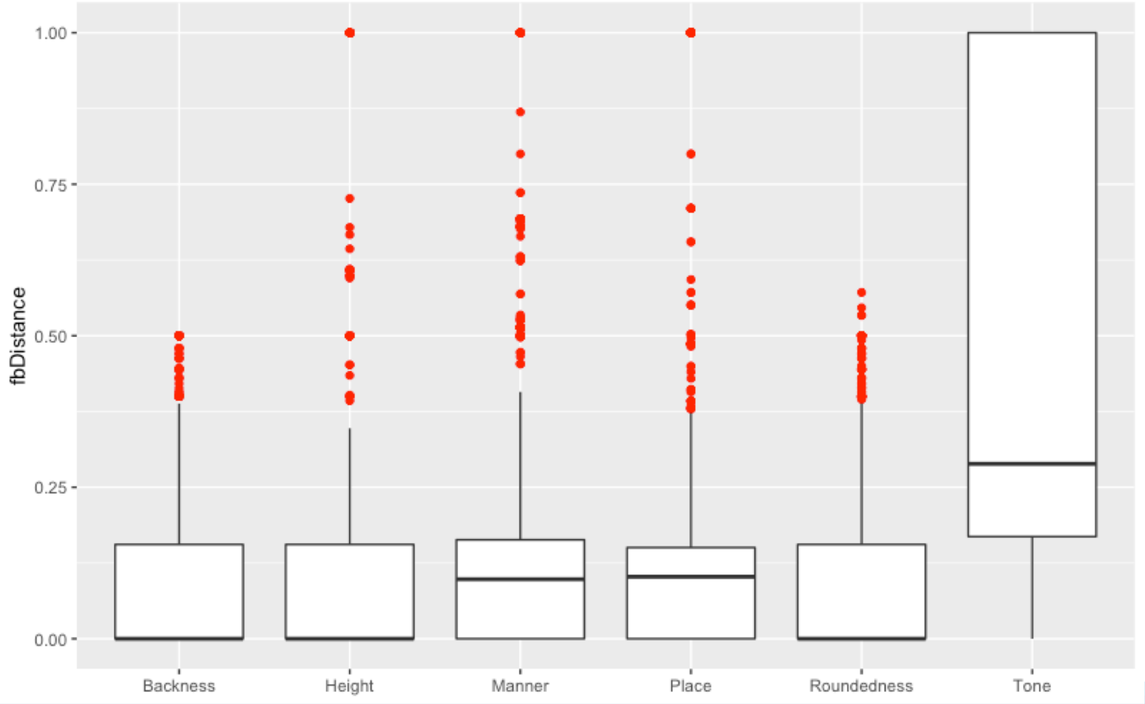


Differences in feature-based Levenshtein distance

- Calculation:

  tone was treated as a dimension and a phoneme type

  the range of values attested in the dataset was mapped onto the range from 0.0 to 1.0 (normalization by overall Levenshtein distance for each cognate pair)



Differences in feature-based Levenshtein distance

- Multinomial regression model:
  - formula: {tone | vowel | consonant} ~ feature-based Levenshtein distance
  - result: highly significant
  - interpretation: in our data set, tone shows more entropy than either vowels or consonants

Summary of results

- in all our measures, tone shows significantly higher entropy/distance
- suggests that it changes faster than segments
- caveats:
  - few, quite closely related varieties
  - lexical entries only
- approach adds transparency and presents the first steps in an empirical assessment of tone volatility

Further Research

- reframing as tonal melody change (Dürr 1987)
- include more varieties from more subgroups, to see if the results change at different time depths
- ideally, replicate the study on the whole Otomanguean family
- even more ideally, replicate the study on other language families (since we'd probably expect there to be cross-linguistic differences)

References

Cahill, Michael. 2011. Tonal diversity in languages of Papua New Guinea. SIL Electronic Working Papers 2011-008.

Campbell, Eric W. (in press). Why is tone change still poorly understood, and how might documentation of less-studied tone languages help? In Patience L. Epps; Na'ama Pat-El & Danny Law (eds.), *Historical linguistics and endangered languages: exploring diversity in language change*. New York: Routledge.

Campbell, Eric W. 2017a. Otomanguean historical linguistics: past, present and prospects for the future. *Language & Linguistics Compass* 11: e12240.

Dürr, Michael. 1987. A preliminary reconstruction of the proto-Mixtec tonal system. *Indiana* 11: 19–61.

Egland, Steven. 1983. La inteligibilidad interdialectal en México: Resultados de algunos sondeos. Mexico, D.F.: Instituto Lingüístico de Verano, A.C. http://www.sil.org/mexico/sondeos/G038b-SondeosInteligibilidad.pdf.

Haudricourt, André-Georges. 1954. De l'origine des tons du vietnamien. *Journal asiatique* 242: 69–82.

Hombert, Jean-Marie, John J. Ohala, & William G. Ewan. 1979. Phonetic explanations for the development of tones. *Language* 55(1): 37–58.

References

INALI. 2009. Catálogo de las lenguas indígenas nacionales: variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas. México: Instituto Nacional de Lenguas Indígenas.

Janda, Richard D. & Brian D. Joseph. 2003. On language, change, and language change—or, of history, linguistics, and historical linguistics. In *The Handbook of historical linguistics*, ed. by Brian D. Joseph & Richard D. Janda, pp. 3–180. Oxford: Blackwell.

Josserand, Judy Kathryn. 1983. Mixtec dialect history. PhD dissertation, Tulane University.

Kingston, John. 2011. Tonogenesis. In *Blackwell Companion to Phonology*, Vol. 4, ed. by Marc van Oostendorp, Colin. J. Ewen, Elizabeth Hume & Keren Rice, pp. 2304–2334. Oxford: Blackwell.

Lewis, M. Paul, Gary F. Simons, & Charles D. Fennig (eds.). 2016. *Ethnologue: Languages of the World, Nineteenth edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Nichols, Johanna. 1995. Diachronically stable structural features. In Henning Andersen (ed.), *Historical linguistics 1993: selected papers from the 11th International Conference on Historical Linguistics, Los Angeles, 16–20 August 1993*, 337–355. Amsterdam: John Benjamins.

References

Ratliff, Martha. 2015. Tonoexodus, tonogenesis, and tone change. In Patrick Honeybone & Joseph Salmons (eds.), *The Oxford Handbook of Historical Phonology*, 245–261. Oxford: Oxford University Press.

Thurgood, Graham. 2002. Vietnamese tonogenesis: revising the model and the analysis. *Diachronica* 19(2): 333–363.

Welmers, Wm. E. 1959. Tonemics, morphotonemics, and tonal morphemes. *General Linguistics* 5(1): 1–9.

Wichmann Søren, & Eric W. Holman. 2009. *Temporal stability for linguistic typological features*. München: Lincom Europa.

Yip, Moira. 2002. *Tone*. Cambridge: Cambridge University Press.

Simple Entropy: Multinomial Regression Output

```
Call:
glm(formula = IS_T ~ ENTROPY, family = "binomial", data = discEntr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6954  -1.2597   0.7366   0.8816   1.2901

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.16602    0.04408   26.45   <2e-16 ***
ENTROPY     -1.42713    0.09638  -14.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5675.5  on 4527  degrees of freedom
Residual deviance: 5451.8  on 4526  degrees of freedom
AIC: 5455.8

Number of Fisher Scoring iterations: 4
```

Anova:

```
# weights:  6 (2 variable)
initial  value 4974.516443
final  value 4968.265832
converged
Analysis of Deviance Table (Type II tests)

Response: SLOT_TYPE
        LR Chisq Df Pr(>Chisq)
ENTROPY   297.52  2  < 2.2e-16 ***
---
```

Simple Entropy: Logistic Regression Output

```
Call:
glm(formula = IS_T ~ ENTROPY, family = "binomial", data = discEntr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6954  -1.2597   0.7366   0.8816   1.2901

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.16602    0.04408   26.45   <2e-16 ***
ENTROPY     -1.42713    0.09638  -14.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5675.5  on 4527  degrees of freedom
Residual deviance: 5451.8  on 4526  degrees of freedom
AIC: 5455.8

Number of Fisher Scoring iterations: 4
```

Feature-based Entropy: Multinomial Regression Output

```
multinom(formula = SLOT_TYPE ~ fbENTROPY, data = fentr.t)

Coefficients:
               (Intercept) fbENTROPY
H_CManner         1.052380 -2.063901
H_CPlace          1.019314 -1.932845
H_CVoicing        1.181418 -2.706678
H_VBackness       1.086304 -2.846573
H_VHeight         1.012230 -2.415687
H_VNasality       1.119236 -3.083280
H_VRoundedness    1.090498 -2.874824

Std. Errors:
               (Intercept) fbENTROPY
H_CManner       0.07696067 0.1228736
H_CPlace        0.07713071 0.1199468
H_CVoicing      0.07623196 0.1391023
H_VBackness     0.07732684 0.1486831
H_VHeight       0.07784013 0.1360475
H_VNasality     0.07708859 0.1565698
H_VRoundedness  0.07729684 0.1495862

Residual Deviance: 24264.23
AIC: 24292.23
```

ANOVA

```
# weights:  16 (7 variable)
initial  value 12622.210158
final  value 12612.294090
converged
Analysis of Deviance Table (Type II tests)

Response: SLOT_TYPE
          LR Chisq Df Pr(>Chisq)
fbENTROPY   960.36  7  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Simple Leveshtein Distance: Multinomial Regression Output

```
multinom(formula = SLOT ~ LDistance, data = gathCl)

Coefficients:
  (Intercept) LDistance
C   0.6623316 -1.497971
V   0.9695221 -2.594826

Std. Errors:
  (Intercept)  LDistance
C  0.05114819 0.08930281
V  0.04985391 0.10021275

Residual Deviance: 12305.93
AIC: 12313.93
```

Click to add text

ANOVA

```
# weights:  6 (2 variable)
initial  value 6555.419526
final  value 6555.419526
converged
Analysis of Deviance Table (Type II tests)

Response: SLOT
          LR Chisq Df Pr(>Chisq)
LDistance   804.91  2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Feature-based Levenshtein Dist: Multinomial Regression Output

```
Call:
multinom(formula = SLOT ~ fbDistance, data = lev.grads.t)

Coefficients:
                 (Intercept) fbDistance
Backness.Grad       1.165194  -5.537687
Height.Grad         1.062135  -4.470778
Manner.Grad         1.029861  -4.181807
Place.Grad          1.009052  -4.005703
Roundedness.Grad    1.150538  -5.370985

Std. Errors:
                 (Intercept) fbDistance
Backness.Grad     0.04673050  0.2026116
Height.Grad       0.04624439  0.1657496
Manner.Grad       0.04609905  0.1559125
Place.Grad        0.04601263  0.1500539
Roundedness.Grad  0.04666508  0.1969311

Residual Deviance: 39582.24
AIC: 39602.24
```

ANOVA

```
# weights:  12 (5 variable)
initial  value 21189.347483
final  value 21189.347483
converged
Analysis of Deviance Table (Type II tests)

Response: SLOT
           LR Chisq Df Pr(>Chisq)
fbDistance   2796.5  5  < 2.2e-16 ***
---
```