Predicting Vowel Harmony class from PMI-score

Lili Szabó

May 18, 2012

Research question

- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?

Research question

- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?
- comparing Dutch and Hungarian

Research question

- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?
- comparing Dutch and Hungarian
  - Dutch: no vowel harmony

Research question

- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?
- comparing Dutch and Hungarian
  - Dutch: no vowel harmony
  - Hungarian: exhibits vowel harmony

Research question

- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?
- comparing Dutch and Hungarian
  - Dutch: no vowel harmony
  - Hungarian: exhibits vowel harmony

Research question

- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?
- comparing Dutch and Hungarian
  - Dutch: no vowel harmony
  - Hungarian: exhibits vowel harmony
- Why is this a relevant question?

Research question

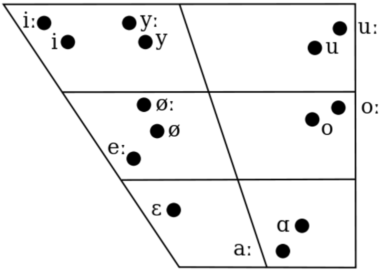
- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?
- comparing Dutch and Hungarian
  - Dutch: no vowel harmony
  - Hungarian: exhibits vowel harmony
- Why is this a relevant question?
  - when learning a language:

Research question

- Does the distribution of vowels differ within and beyond word boundaries in a language with vowel harmony?
- comparing Dutch and Hungarian
  - Dutch: no vowel harmony
  - Hungarian: exhibits vowel harmony
- Why is this a relevant question?
  - when learning a language:
  - does vowel harmony help with word segmentation?

Hungarian vowels

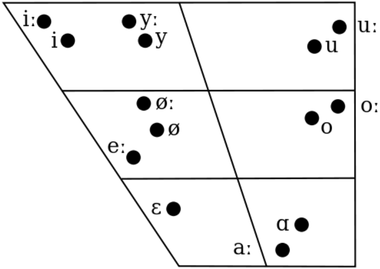
Figure: Hungarian Vowel Chart



- backness feature of vowels

Hungarian vowels

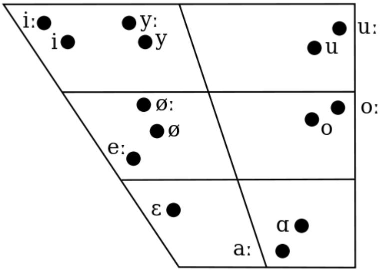
Figure: Hungarian Vowel Chart



- backness feature of vowels
- vowels within words agree in their backness feature

Hungarian vowels

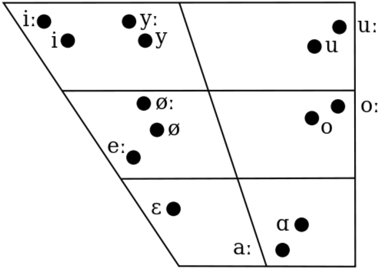
Figure: Hungarian Vowel Chart



- backness feature of vowels
- vowels within words agree in their backness feature
- important role in suffixation

Hungarian vowels

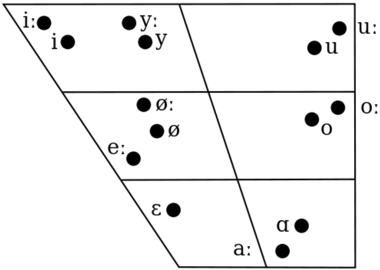
Figure: Hungarian Vowel Chart



- backness feature of vowels
- vowels within words agree in their backness feature
- important role in suffixation
- neutral vowels: *e, é, i, í*

Hungarian vowels

Figure: Hungarian Vowel Chart



- backness feature of vowels
- vowels within words agree in their backness feature
- important role in suffixation
- neutral vowels: *e, é, i, í*
- orthography of the 14 Hungarian vowels is completely phonetic

Dative suffix - nAk (nak/nek)

- remek alma
  wonderfulNOM appleNOM
  'wonderful apple'

Dative suffix - nAk (nak/nek)

- remek alma
  wonderfulNOM appleNOM
  'wonderful apple'
- almának
  appleDAT
  'to the apple'

Dative suffix - nAk (nak/nek)

- remek alma
  wonderfulNOM appleNOM
  'wonderful apple'
- almának
  appleDAT
  'to the apple'
- remeknek
  wonderfulDAT
  'to the wonderful'
- Marinak
  MaryDAT
  'to Mary'

Dutch vowels and diphthongs

- 13 monophthongs, 4 diphthongs

Dutch vowels and diphthongs

- 13 monophthongs, 4 diphthongs
- 25 orthographic symbols

Dutch vowels and diphthongs

- 13 monophthongs, 4 diphthongs
- 25 orthographic symbols
- orthography is not entirely phonetic:

Dutch vowels and diphthongs

- 13 monophthongs, 4 diphthongs
- 25 orthographic symbols
- orthography is not entirely phonetic:
  - e.g. pronunciation of *ij* and *ei* is identical in *bijt* and *ei*

Dutch vowels and diphthongs

- 13 monophthongs, 4 diphthongs
- 25 orthographic symbols
- orthography is not entirely phonetic:
  - e.g. pronunciation of *ij* and *ei* is identical in *bijt* and *ei*
  - but in this project they were treated as separate symbols in the transcription

Corpora from CHILDES - child directed speech

Table: Corpora details

language	Dutch	Hungarian
token	749755	93254
type	16002	9259
type-token ratio	0.021	0.099
avg. vowels / word	1.283953	1.675671

Pointwise Mutual Information

- a measure to score associations (e.g. collocations)
- how two events co-occur

Pointwise Mutual Information

- a measure to score associations (e.g. collocations)
- how two events co-occur
- comparing expected vs. observed probabilities
- $\frac{observed}{expected}$ co-occurances

Pointwise Mutual Information

- a measure to score associations (e.g. collocations)
- how two events co-occur
- comparing expected vs. observed probabilities
- $\frac{observed}{expected}$ co-occurances
- are 2 vowels in consecutive syllables within a word co-occuring more often than it would be expected from their frequency in the data?

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$
- range: -inf;+inf

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$
- range: -inf;+inf
  - 0: as often as expected (log(1))

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$
- range: -inf;+inf
  - 0: as often as expected (log(1))
  - negatives values: less than expected

Pointwise Mutual Information - properties

- $\frac{observed}{expected}$ co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$
- range: -inf;+inf
  - 0: as often as expected (log(1))
  - negatives values: less than expected
  - positive values: more than expected

Pointwise Mutual Information - calculation

- after preprocessing the data

Pointwise Mutual Information - calculation

- after preprocessing the data
- using Python-NLTK to calculate scores

Pointwise Mutual Information - calculation

- after preprocessing the data
- using Python-NLTK to calculate scores

Pointwise Mutual Information - calculation

- after preprocessing the data
- using Python-NLTK to calculate scores
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$

Pointwise Mutual Information - calculation

- after preprocessing the data
- using Python-NLTK to calculate scores
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$
- Dutch: PMI(a,e) = -0.03
- Hungarian: PMI(a,e) = -4.3

Pointwise Mutual Information - calculation

- after preprocessing the data
- using Python-NLTK to calculate scores
- $\log_2 \frac{p(a,e)}{p(a)p(e)}$
- Dutch: PMI(a,e) = -0.03
- Hungarian: PMI(a,e) = -4.3
- Dutch: PMI(e,e) = -0.09
- Hungarian: PMI(e,e) = 0.35

Smoothing

- add-one smoothing (adding all possible vowel pairs with count 1)

Smoothing

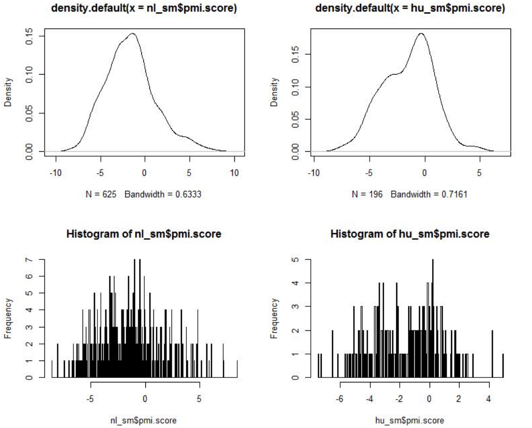
- add-one smoothing (adding all possible vowel pairs with count 1)
- PMI-scores of these bigrams range from lowest to highest

Smoothing

- add-one smoothing (adding all possible vowel pairs with count 1)
- PMI-scores of these bigrams range from lowest to highest
- low frequency effect for rare Dutch diphthongs

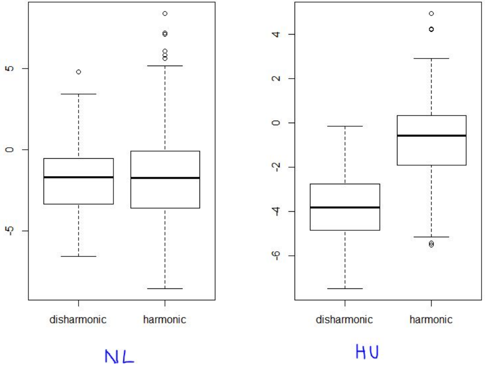
Distribution

Figure: Distribution of PMI-scores



Boxplot

Figure: PMI-scores wrt Harmony-class



Why to use logistic regression?

- to predict VH-class from PMI-score

Why to use logistic regression?

- to predict VH-class from PMI-score
- are the log odds $\log \frac{p}{1-p}$ of harmony class predictable from PMI-score?

Probability, odds, log odds - range

- p: [0:1]

Probability, odds, log odds - range

- p: [0:1]
- odds: $\frac{p}{1-p}$ = [0:+inf]

Probability, odds, log odds - range

- p: [0:1]
- odds: $\frac{p}{1-p}$ = [0:+inf]
- log odds: [-inf:+inf]

Variables

- independent: PMI-score - numeric

Variables

- independent: PMI-score - numeric
- dependent: VH-class - binary (harmonic vs. disharmonic)

Variables

- independent: PMI-score - numeric
- dependent: VH-class - binary (harmonic vs. disharmonic)
- simple model (1 independent variable)

Assumptions

- no normality of independent variable and residuals is required

Assumptions

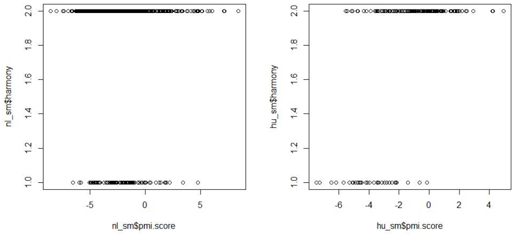
- no normality of independent variable and residuals is required
- dependent variable is dichotomic: true
- (information loss: neutral class in harmony)

Assumptions

- no normality of independent variable and residuals is required
- dependent variable is dichotomic: true
- (information loss: neutral class in harmony)
- independent variables are linearly related to the log odds

PMI-score and harmony-class

Figure: PMI-score and harmony-class



Intercept model


```
> with(hu, table(harmony))
harmony
  0   1
  48 148
> 148 / (48+148) # prob of harmonic pairs
[1] 0.755102
> hu_m1 = glm(formula = hu$harmony ~ 1, family = binomial(link = "logit")
> summary(hu_m1) # intercept model
[...]
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5805     0.2075   7.616 2.62e-14 ***
[...]
> antilogit <- function(x) { exp(x) / (1 + exp(x) ) } # logit to prob
> antilogit(1.5805)
[1] 0.8292753
```

Logit model


```
> hu_m2 = glm(formula = hu$harmony ~ hu$pmi.score, family = binomial(link
> summary(hu_m2) # model with pmi.score
[...]
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.2192     0.6661   6.334 2.38e-10 ***
hu$pmi.score   1.0721     0.1887   5.682 1.33e-08 ***
[...]
  > anova(hu_m1, hu_m2)
Analysis of Deviance Table

Model 1: hu$harmony ~ 1
Model 2: hu$harmony ~ hu$pmi.score
  Resid. Df Resid. Dev Df Deviance
1       163    149.911
2       162     81.983  1   67.927
```

Coefficients - how to interpret odds ratios?


```
> 1-pchisq(67.927,df=1) #computing the chi-sqare probability of deviance
[1] 2.220446e-16
> exp(hu_m2$coefficients)
  (Intercept) hu$pmi.score
    67.979974     2.921596
```

- one unit increase in pmi.score, the odds of being a harmonic pair (versus not being harmonic) increase by a factor of 2.92

Logit models of Dutch

- intercept significant: p = 2e-16

Logit models of Dutch

- intercept significant: p = 2e-16
- pmi.score: non-significant, p = 0.71

Logit models of Dutch

- intercept significant: $p = 2e\text{-}16$
- pmi.score: non-significant, $p = 0.71$
- pmi.score in the without smoothing data is not significant either: $p = 0.308$

Conclusion

- PMI-score of of vowel pairs (vowels in neighbouring syllables)

Conclusion

- PMI-score of of vowel pairs (vowels in neighbouring syllables)
    - is NOT a predictor of VH-class of Dutch vowel pairs

Conclusion

- PMI-score of of vowel pairs (vowels in neighbouring syllables)
  - is NOT a predictor of VH-class of Dutch vowel pairs
  - is a predictor of VH-class of Hungarian vowel pairs