Predicting Vowel Harmony class from PMI-score

Lili Szabó

May 18, 2012
Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

1. Outline
2. Introduction
3. Vowel harmony
4. Dutch vowels - diphthongs
5. Data
6. PMI
7. Results - PMI
8. Logistic regression
   - Comparing Hungarian and Dutch
   - Hungarian data analysis
   - Dutch data analysis
9. Discussion
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

- backness feature of vowels
- vowels within words agree in their backness feature
- important role in suffixation
Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline
Introduction
Vowel harmony
Dutch vowels - diphthongs
Data
PMI
Results - PMI
Logistic regression
Comparing Hungarian and Dutch
Hungarian data analysis
Dutch data analysis
Discussion

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

- 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
  wonderful NOM apple NOM
  'wonderful apple'

- **almának**
  apple DAT
  '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
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

<table>
<thead>
<tr>
<th>language</th>
<th>Dutch</th>
<th>Hungarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>token</td>
<td>749755</td>
<td>93254</td>
</tr>
<tr>
<td>type</td>
<td>16002</td>
<td>9259</td>
</tr>
<tr>
<td>type-token ratio</td>
<td>0.021</td>
<td>0.099</td>
</tr>
<tr>
<td>avg. vowels / word</td>
<td>1.283953</td>
<td>1.675671</td>
</tr>
</tbody>
</table>
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
  \[
  \frac{\text{observed}}{\text{expected}} \text{ co-occurrences}
  \]
Pointwise Mutual Information

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

- \( \frac{\text{observed}}{\text{expected}} \) co-occurrences

- \( \log_2 \left( \frac{\text{observed}}{\text{expected}} \right) \)

- range: \(-\infty;+\infty\)
  - 0: as often as expected (log(1))
  - negative values: less than expected
  - positive values: more than expected

- symmetric: \( \text{pmi}(x,y) = \text{pmi}(y,x) \)
- high score for low frequency items
Pointwise Mutual Information - properties

- \( \frac{\text{observed}}{\text{expected}} \) co-occurrences
- derived from Mutual Information

Pointwise Mutual Information (PMI) is a measure of the strength of dependence between two events. It is defined as the logarithm (usually base 2) of the ratio of the observed co-occurrences to the expected co-occurrences. Mathematically, it can be expressed as:

\[
\text{PMI}(X, Y) = \log_2 \left( \frac{p(a, e)}{p(a)p(e)} \right)
\]

- **symmetric**: PMI(x,y) = PMI(y,x)
- **high score for low frequency items**
- **range**: \(-\infty; +\infty\)
  - 0: as often as expected (\(\log(1)\))
  - negatives values: less than expected
  - positive values: more than expected
Pointwise Mutual Information - properties

- **observed** co-occurrences
- **expected**

derived from Mutual Information

symmetric: \( \text{pmi}(x,y) = \text{pmi}(y,x) \)
Pointwise Mutual Information - properties

- **observed** co-occurrences compared to **expected**
- derived from Mutual Information
- symmetric: $\text{pmi}(x,y) = \text{pmi}(y,x)$
- high score for low frequency items
Pointwise Mutual Information - properties

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

- observed co-occurrences
- derived from Mutual Information
- symmetric: $\text{pmi}(x,y) = \text{pmi}(y,x)$
- high score for low frequency items
  \[
  \log_2 \frac{p(a,e)}{p(a)p(e)}
  \]
- range: $-\infty;+\infty$
Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline
Introduction
Vowel harmony
Dutch vowels - diphthongs
Data
PMI
Results - PMI
Logistic regression
Comparing Hungarian and Dutch
Hungarian data analysis
Dutch data analysis
Discussion

Pointwise Mutual Information - properties

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

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

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

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels - diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian and Dutch

Hungarian data analysis

Dutch data analysis

Discussion

Pointwise Mutual Information - calculation

- after preprocessing the data
Pointwise Mutual Information - calculation

- after preprocessing the data
- using Python-NLTK to calculate scores
Predicting Vowel Harmony class from PMI-score
Lili Szabó

Outline
Introduction
Vowel harmony
Dutch vowels - diphthongs
Data
PMI
Results - PMI
Logistic regression
Comparing Hungarian and Dutch
Hungarian data analysis
Dutch data analysis
Discussion

Pointwise Mutual Information - calculation

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

Pointwise Mutual Information - calculation

Pointwise Mutual Information - calculation
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
Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline
Introduction
Vowel harmony
Dutch vowels - diphthongs
Data
PMI
Results - PMI
Logistic regression
Comparing Hungarian and Dutch
Hungarian data analysis
Dutch data analysis
Discussion

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: \( PMI(a, e) = -4.3 \)
Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline
- Introduction
- Vowel harmony
- Dutch vowels - diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian and Dutch

Hungarian data analysis

Dutch data analysis

Discussion

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
Figure: Distribution of PMI-scores

![Distribution of PMI-scores](image_url)
Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline
Introduction
Vowel harmony
Dutch vowels - diphthongs
Data
PMI
Results - PMI
Logistic regression
Comparing Hungarian and Dutch
Hungarian data analysis
Dutch data analysis
Discussion

Figure: PMI-scores wrt Harmony-class

Boxplot
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

- \( p: [0:1] \)
- \( \text{odds}: \frac{p}{1-p} = [0:+\infty] \)
Probability, odds, log odds - range

- $p$: [0:1]
- odds: $\frac{p}{1 - p} = [0:+\infty]$  
- log odds: $[-\infty:+\infty]$
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
Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline
Introduction
Vowel harmony
Dutch vowels - diphthongs
Data
PMI
Results - PMI
Logistic regression
Comparing Hungarian and Dutch
Hungarian data analysis
Dutch data analysis
Discussion

Figure: PMI-score and harmony-class
Predicting Vowel Harmony class from PMI-score
Lili Szabó

Outline
Introduction
Vowel harmony
Dutch vowels - diphthongs
Data
PMI
Results - PMI
Logistic regression
Comparing Hungarian and Dutch
Hungarian data analysis
Dutch data analysis
Discussion

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 = "logit"))
> 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: hu$harmony ~ 1</td>
<td>163</td>
<td>149.911</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2: hu$harmony ~ hu$pmi.score</td>
<td>162</td>
<td>81.983</td>
<td>1</td>
<td>67.927</td>
</tr>
</tbody>
</table>
Coefficients - how to interpret odds ratios?

> 1-pchisq(67.927, df=1) #computing the chi-square 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: $p = 2e-16$
- 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 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 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