Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

Discussion

# Predicting Vowel Harmony class from PMI-score

Lili Szabó

May 18, 2012

#### Predicting Vowel Harmony class from PMI-score

### Outline

### Outline



3 Vowel harmony

4 Dutch vowels - diphthongs

Data

6 PMI

Results - PMI

### 8 Logistic regression

• Comparing Hungarian and Dutch

- Hungarian data analysis
- Dutch data analysis



#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

#### Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

Discussion

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

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

#### Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

analysis Dutch data analysis

Discussion

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

• comparing Dutch and Hungarian

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

#### Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

analysis Dutch data analysis

Discussion

• 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

#### 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

• 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

#### 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

• 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

#### 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

analysis Dutch data analysis

Discussion

• 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?

#### 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 dat

analysis Dutch data

Discussion

• 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:

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

### Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data

analysis

Discussion

- 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?

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

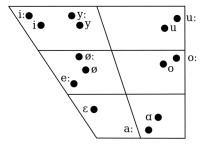
Comparing Hungarian an Dutch

Hungarian data analysis Dutch data

analysis

Discussion

Figure: Hungarian Vowel Chart



backness feature of vowels

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

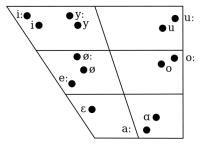
Hungarian and Dutch

Hungarian data analysis Dutch data

analysis

Discussion

Figure: Hungarian Vowel Chart



◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

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

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

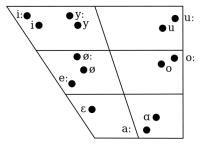
Logistic regression

Hungarian and Dutch

Hungarian data analysis Dutch data

allalysis

Figure: Hungarian Vowel Chart



- 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

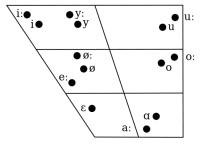
Hungarian and Dutch

Hungarian data analysis Dutch data

analysis

Discussion

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

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

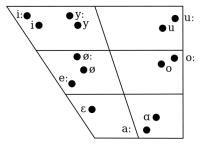
Hungarian and Dutch

Hungarian data analysis Dutch data

analysis

Discussion

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)

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian dat: analysis Dutch data analysis

Discussion

• remek alma wonderfulNOM appleNOM 'wonderful apple'

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

# Dative suffix - nAk (nak/nek)

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

Discussion

• remek alma wonderfulNOM appleNOM 'wonderful apple'

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

 almának appleDAT 'to the apple'

# Dative suffix - nAk (nak/nek)

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian ar Dutch Hungarian da

Dutch data analysis

Discussion

• remek alma wonderfulNOM appleNOM 'wonderful apple'

- almának appleDAT 'to the apple'
- remeknek wonderfulDAT 'to the wonderful'
- Marinak MaryDAT *'to Mary'*



Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

analysis Dutch data analysis

Discussion

• 13 monophthongs, 4 diphthongs

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

Discussion

• 13 monophthongs, 4 diphthongs

• 25 orthographic symbols

▲ロト ▲園 ▶ ▲ 臣 ▶ ▲ 臣 ▶ 臣 のへ⊙

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

### Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

Discussion

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

#### 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

- 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*

### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

#### Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data

Discussion

- 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



Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

### Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian ar Dutch

Hungarian data analysis Dutch data

analysis

Discussion

### 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

・ロト ・ 理 ト ・ ヨ ト ・ ヨ ト

∃ \0 < \0</p>

### Pointwise Mutual Information

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data

• a measure to score associations (e.g. collocations)

▲ロト ▲周ト ▲ヨト ▲ヨト ヨー のくで

how two events co-occur

# Pointwise Mutual Information

### 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

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

▲ロト ▲帰 ト ▲ ヨ ト ▲ ヨ ト ・ ヨ ・ の Q ()

• <u>expected</u> co-occurances

# Pointwise Mutual Information

### Predicting Vowel Harmony class from PMI-score

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

- a measure to score associations (e.g. collocations)
- how two events co-occur
- comparing expected vs. observed probabilities observed
- $\frac{1}{expected}$  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?

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

### PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data

anarysis

Discussion

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

▲ロト ▲帰 ト ▲ ヨ ト ▲ ヨ ト ・ ヨ ・ の Q ()

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

### PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data

Discussion

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

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

### PMI

Results - PMI

Logistic regression

Hungarian an Dutch

analysis Dutch data

Discussion

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

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

### PMI

Results - PMI

Logistic regression

Hungarian an Dutch

Hungarian data analysis Dutch data analysis

Discussion

- $\frac{1}{expected}$  co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

### 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

observed

- $\frac{1}{expected}$  co-occurances
- derived from Mutual Information
- symmetric: pmi(x,y)=pmi(y,x)
- high score for low frequency items

•  $\log_2 \frac{p(a, e)}{p(a)p(e)}$ 

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

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

- $\frac{expected}{expected}$  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

### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

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

observed

- 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))

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

### 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 dat analysis
- Dutch dat
- Discussion

observed

- 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

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

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

observed

- 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



Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

#### PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian dat: analysis Dutch data analysis

Discussion

### • after preprocessing the data

・ロト・日本・ヨト・ヨー シへの

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

#### PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

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

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

#### PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

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

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmon

Dutch vowels diphthongs

Data

#### PMI

Results - PMI

Logistic regression

Comparing Hungarian and Dutch

Hungarian data analysis Dutch data

analysis

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



▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

#### Predicting Vowel Harmony class from PMI-score

#### Lili Szabó

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

- 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

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

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

- 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

- Predicting Vowel Harmony class from PMI-score
  - Lili Szabó
- Outline
- Introduction
- Vowel harmony
- Dutch vowels diphthongs
- Data
- PМ
- Results PMI
- Logistic regression
- Comparing Hungarian and Dutch
- Hungarian data analysis Dutch data
- \_\_\_\_

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

## Smoothing

- Predicting Vowel Harmony class from PMI-score
  - Lili Szabó
- Outline
- Introduction
- Vowel harmony
- Dutch vowels diphthongs
- Data
- PM
- Results PMI
- Logistic regression
- Comparing Hungarian and Dutch
- Hungarian data analysis Dutch data
- anaiysis
- Discussion

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

• PMI-scores of these bigrams range from lowest to highest

## Smoothing

#### 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

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

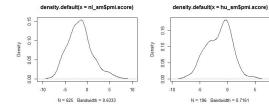
- PMI-scores of these bigrams range from lowest to highest
- low frequency effect for rare Dutch diphthongs

### Distribution

#### Predicting Vowel Harmony class from PMI-score

#### Results - PMI

### Figure: Distribution of PMI-scores



5

Histogram of nl\_sm\$pmi.score

0

nl smSpmi.score

9

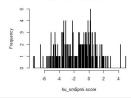
N

-5

Frequency



0



(日)、 э.

### Boxplot

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

#### Results - PMI

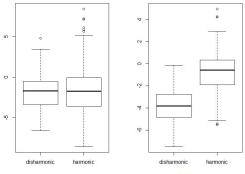
Logistic regression

Comparing Hungarian and Dutch Hungarian dat analysis

Dutch data analysis

Discussion

### Figure: PMI-scores wrt Harmony-class



NL

HU

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへで

### Why to use logistic regression?



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

### • to predict VH-class from PMI-score

## Why to use logistic regression?

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

- 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

Predicting Vowel Harmony class from PMI-score Lili Szabó	
Outline	
Introduction	
Vowel harmony	
Dutch vowels - diphthongs	• p: [0:1]
Data	
РМІ	
Results - PMI	
Logistic regression Comparing Hungarian and Dutch Hungarian data analysis Dutch data analysis	
Discussion	

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

### Probability, odds, log odds - range

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?



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

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

### Probability, odds, log odds - range



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

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

log odds: [-inf:+inf]

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

### Variables

#### 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

### • independent: PMI-score - numeric

<ロ> <畳> < 三> < 三> < 三> < 三> < 三> < 三</p>

### Variables

#### 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

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

### Variables

#### 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

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

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• simple model (1 independent variable)

### Assumptions



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

• no normality of independent variable and residuals is required

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

### Assumptions

#### 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

• no normality of independent variable and residuals is required

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

- dependent variable is dichotomic: true
- (information loss: neutral class in harmony)

### Assumptions

#### 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

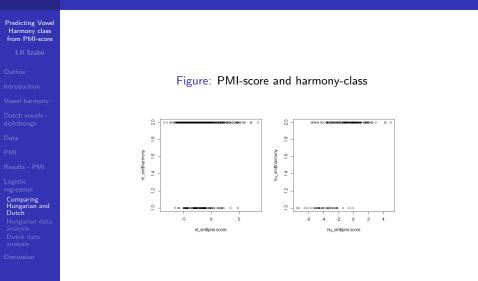
Discussion

• no normality of independent variable and residuals is required

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

- 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



◆□▶ ◆□▶ ◆臣▶ ◆臣▶ ─臣 ─の�?

```
Intercept model
Predicting Vowel
Harmony class
from PMI-score
             > 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
Hungarian data
              > antilogit(1.5805)
analysis
              [1] 0.8292753
```

### Logit model

#### Predicting Vowel Harmony class from PMI-score

#### Lili Szabó

```
Outline
```

Introduction

Vowel harmon

Dutch vowels diphthongs

Data

PMI

Results - PM

Logistic regression

Hungarian an Dutch

Hungarian data analysis Dutch data analysis

Discussion

```
> hu_m2 = qlm(formula = hu$harmony ~ hu$pmi.score, family = binomial(lin)
> summarv(hu m2) # model with pmi.score
[...]
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)
               4.2192
                          0.6661 6.334 2.38e-10 ***
              1.0721 0.1887 5.682 1.33e-08 ***
hu$pmi.score
[...]
 > anova(hu m1, hu m2)
Analysis of Deviance Table
Model 1: hu$harmonv ~ 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
```

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 三臣 - のへ(?)

### Coefficients - how to interpret odds ratios?

```
    Predicting Vowel
Harmony class
from PMI-score
    -

    Lili Scabb
    -

    Outline
    -

    Introduction
    -

    Vowel harmony
    > 1-pchisq(67.927, df=1) #computing the chi-sqare probability of deviance
[1] 2.220446e-16

    Dutch vowels -
diphthongs
    -

    Data
    67.979974

    PMI
    -

    Results - PMI
    -
```

- Logistic regression
- Comparing Hungarian and Dutch
- Hungarian data analysis Dutch data
- analysis
- Discussion

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

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

### Logit models of Dutch

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 significant: p = 2e-16

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

### Logit models of Dutch

Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis

Dutch data analysis

Discussion

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

▲ロト ▲周ト ▲ヨト ▲ヨト ヨー のくで

### Logit models of Dutch

#### 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 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

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

### Conclusion

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

analysis Dutch data analysis

Discussion

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

### Conclusion

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

analysis Dutch data analysis

Discussion

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

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

### Conclusion

#### Predicting Vowel Harmony class from PMI-score

Lili Szabó

Outline

Introduction

Vowel harmony

Dutch vowels diphthongs

Data

PMI

Results - PMI

Logistic regression

Comparing Hungarian an Dutch

Hungarian data analysis Dutch data analysis

#### Discussion

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

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• is NOT a predictor of VH-class of Dutch vowel pairs

• is a predictor of VH-class of Hungarian vowel pairs