A Sociolinguistic Analysis of Linguistically Sensitive Dialectal Word Pronunciation Distances

Martijn Wieling, University of Groningen

Computational approaches to the study of dialectal and typological variation, ESSLI 2012, August 6 - 10, 2012

Overview

- Linguistically sensitive segment distances
  - Why use sensitive segment distances?
  - Obtaining sensitive segment distances
  - Evaluating the quality of sensitive segment distances
- Sociolinguistic factors influencing Dutch dialect distances
  - The Dutch dialect dataset
  - Modeling the effect of geography
  - Mixed-effects regression modeling
  - Important predictors

Collaborators



The need for sensitive segment distances (1)

- In our research on language variation, we employ pronunciation distances (on the basis of alignments)
- We would like to improve alignment quality and the distances

- There is no widely accepted procedure to determine phonetic similarity (Laver, 1994)
- Here we use the distribution of pronunciation variation to determine similarity
- In line with language as “un systême oû tout se tient” (focus on relations between items, not items themselves; Meillet, 1903)

The need for sensitive segment distances (2)

- We evaluate the phonetic sound distances we automatically obtain by comparing them to acoustic (vowel) distances
- In an earlier study (Wieling, Prokić and Nerbonne, 2009), we already showed that the method improves alignment quality significantly

Our starting point: the Levenshtein distance

Restriction: vowels are not aligned with consonants

- The Levenshtein distance measures the minimum number of insertions, deletions and substitutions to transform one string into another

mɔəlkə	delete ɔ	1
məlkə	subst. ə/ɛ	1
mɛlkə	delete ə	1
mɛlk	insert ə	1
mɛlək		
		4

m	ɔ	ə	l		k	ə
m		ɛ	l	ə	k	
	1	1		1		1

- Note that the alignment results in an implicit identification of sound correspondences

Our starting point: the Levenshtein distance

Restriction: vowels are not aligned with consonants

- The Levenshtein distance measures the minimum number of insertions, deletions and substitutions to transform one string into another

mɔəlkə	delete ɔ	1
məlkə	subst. ə/ɛ	1
mɛlkə	delete ə	1
mɛlk	insert ə	1
mɛlək		
		4

m	ɔ	ə	l		k	ə
m		ɛ	l	ə	k	
	1	1		1		1

- Note that the alignment results in an implicit identification of sound correspondences

Counting sound segment correspondences

- Counting the frequency of sound segments (in the Levenshtein alignments)

p	b	...	ʊ	u	Total
5×10^5	2×10^5	...	90,000	9×10^5	10^8

- Counting the frequency of the aligned sound segments (in the Levenshtein alignments)

	p	b	...	ʊ	u	
p	2×10^5	10,650	...	0	0	
b		88,000	...	0	0	
⋮			⋮	⋮	⋮	
ʊ				65,400	5,500	
u					4×10^5	
						Total: 5×10^7

- Probability of observing [p]: 5×10^5 / 10^8 = 0.005 (0.5%)
- Probability of observing [b]: 2×10^5 / 10^8 = 0.002 (0.2%)
- Probability of observing [p]:[b]: 10,650 / 5×10^7 = 0.0002 (0.02%)

Association strength between segment pairs

- Pointwise Mutual Information (PMI): assesses degree of statistical dependence between aligned segments (x and y)

$$\mathrm{PMI}(x, y) = \log_2 \left(\frac{p(x, y)}{p(x)\, p(y)} \right)$$

- $p(x, y)$: relative occurrence of the aligned segments x and y in the whole dataset
- $p(x)$ and $p(y)$: relative occurrence of x and y in the whole dataset
- The greater the PMI value, the more segments tend to cooccur in correspondences

Association strength between segment pairs

- Probability of observing [p]:[b]: 10,650 / 5×10^7 = 0.0002
- Probability of observing [p]: 5×10^5 / 10^8 = 0.005
- Probability of observing [b]: 2×10^5 / 10^8 = 0.002

$$\text{PMI}(x, y) = \log_2 \left(\frac{p(x, y)}{p(x)\, p(y)} \right) \Rightarrow$$

$$\text{PMI}([\text{p}], [\text{b}]) = \log_2 \left(\frac{0.0002}{0.005 \times 0.002} \right)$$

$$\text{PMI}([\text{p}], [\text{b}]) \approx 4.3$$

Using PMI values with the Levenshtein algorithm

- Idea: use association strength to weight edit operations
- PMI is large for strong associations, so invert it (0 - PMI)
  - Strongly associated segments will have a low distance
- PMI range varies, so normalize it between 0 and 1.
- Use PMI-induced weights as costs in Levenshtein algorithm
  - Cost of substituting identical sound segments is always set to 0

The PMI-based Levenshtein algorithm

- We use the standard Levenshtein algorithm to calculate the initial PMI weights and convert these to costs (i.e. sound distances)
- These sensitive sound distances are then used as edit operation costs in the Levenshtein algorithm to obtain new alignments, new counts, and new PMI sound distances
- This process is repeated until alignments and PMI sound distances stabilize
- Besides new alignments, this procedure automatically yields sensitive sound segment distances

m	ɔ		ə	l		k	ə	
m			ɛ	l	ə	k		
	0.20		0.15		0.12		0.12	

The PMI-based Levenshtein algorithm

- We use the standard Levenshtein algorithm to calculate the initial PMI weights and convert these to costs (i.e. sound distances)
- These sensitive sound distances are then used as edit operation costs in the Levenshtein algorithm to obtain new alignments, new counts, and new PMI sound distances
- This process is repeated until alignments and PMI sound distances stabilize
- Besides new alignments, this procedure automatically yields sensitive sound segment distances

m	ɔ	ə	l		k	ə
m		ɛ	l	ə	k	
	0.20	0.15		0.12		0.12

Pronunciation data

- Six independent dialect data sets (IPA pronunciations)
  - Dutch: 562 words in 613 locations (Wieling et al., 2007)
  - German: 201 words in 186 locations (Nerbonne and Siedle, 2005)
  - U.S. English: 153 words in 483 locations (Kretzschmar, 1994)
  - Bantu (Gabon): 160 words in 53 locations (Alewijnse et al., 2007)
  - Bulgarian: 152 words in 197 locations (Prokić et al., 2009)
  - Tuscan: 444 words in 213 locations (Montemagni et al., in press)
- For all datasets sound segment distances are obtained using the PMI-based Levenshtein algorithm
  - We use a slightly adapted version: ignoring identical sound segment substitutions in the counts

Acoustic data

- For the evaluation, we obtained acoustic vowel measurements (F1 and F2) reported in the scientific literature
  - Pols et al. (1973; NL), van Nierop et al. (1973; NL), Sendlmeier and Seebode (2006; GER), Hillenbrand et al. (1995; US), Nurse and Phillipson (2003, p. 22; BAN), Lehiste and Popov (1970; BUL), Calamai (2003; TUS)
- To determine acoustic vowel distance, we calculate the Euclidean distance of the formant frequencies
  - Our perception of frequency is non-linear and calculating the Euclidean distance on the basis of Hertz values would not give enough weight to the first formant
  - We therefore first scale the Hertz frequencies to Bark

Comparison procedure between acoustic and PMI distances

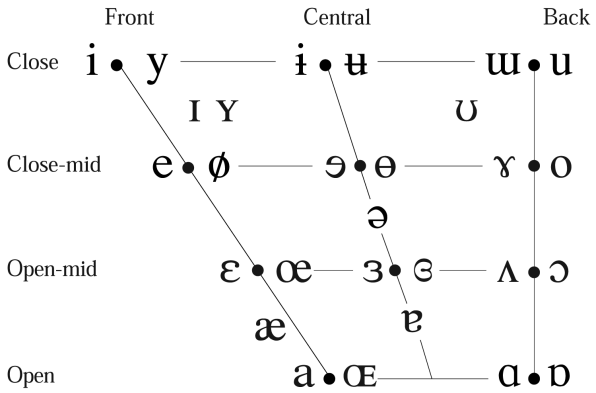
- We assess the relation between the generated and acoustic distances using the Pearson correlation
- We visualize the relative position of the sound segments by applying multidimensional scaling (MDS) to the distance matrices
  - Missing distances are not allowed in the (classical) MDS procedure, so in some cases not all sound segments are visualized

Acoustic vs. PMI vowel distances

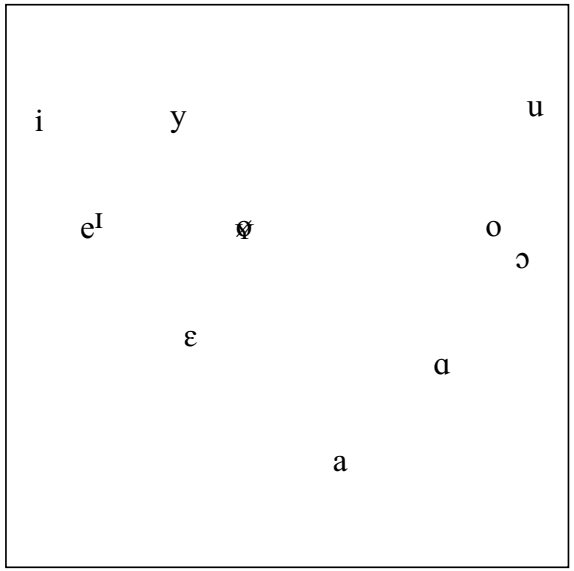
	Pearson's r	Explained variance (r^2)
Dutch	0.672	45.2%
Dutch w/o Frisian	0.686	47.1%
German	0.633	40.1%
German w/o ə	0.785	61.6%
US English	0.608	37.0%
Bantu	0.642	41.2%
Bulgarian	0.677	45.8%
Tuscan	0.758	57.5%

MDS visualization of Dutch vowels

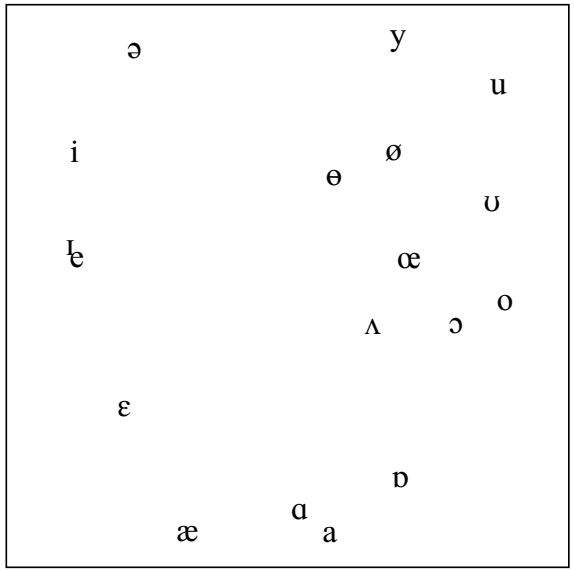
PMI visualization captures 76% of the variation



(a) IPA



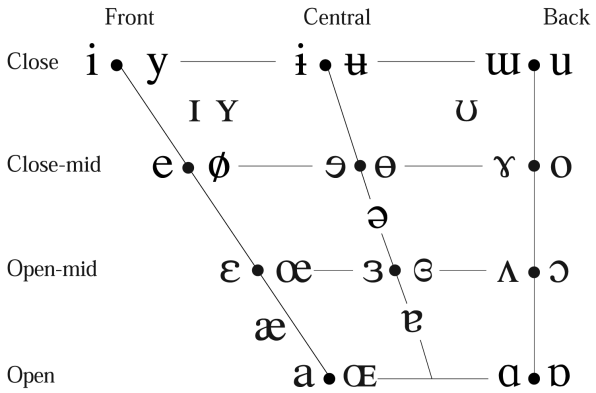
(b) Acoustics



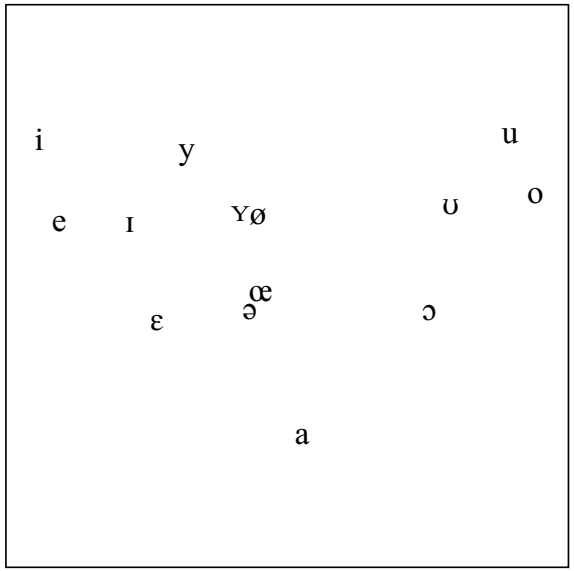
(c) PMI distances

MDS visualization of German vowels

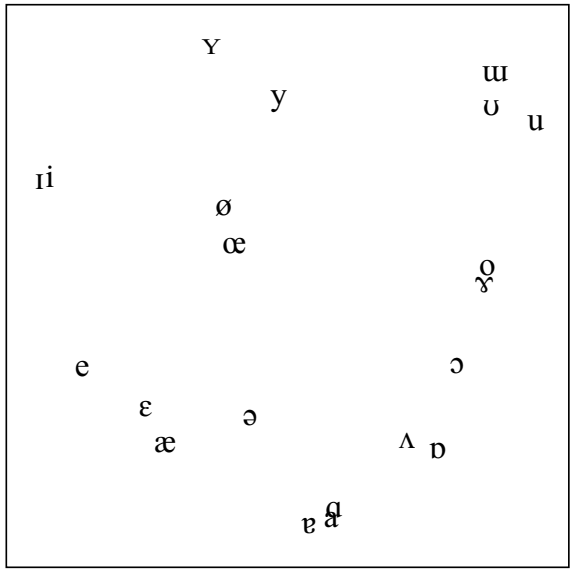
PMI visualization captures 70% of the variation



(a) IPA



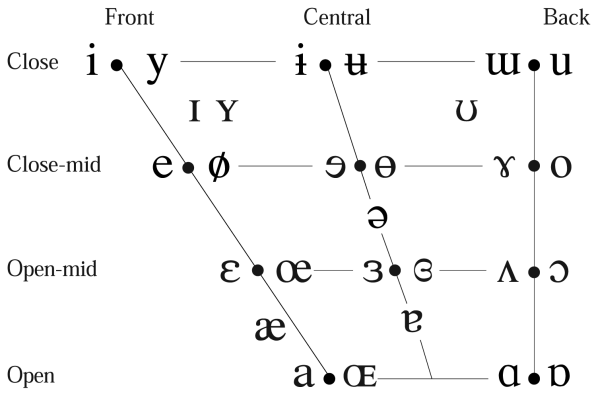
(b) Acoustics



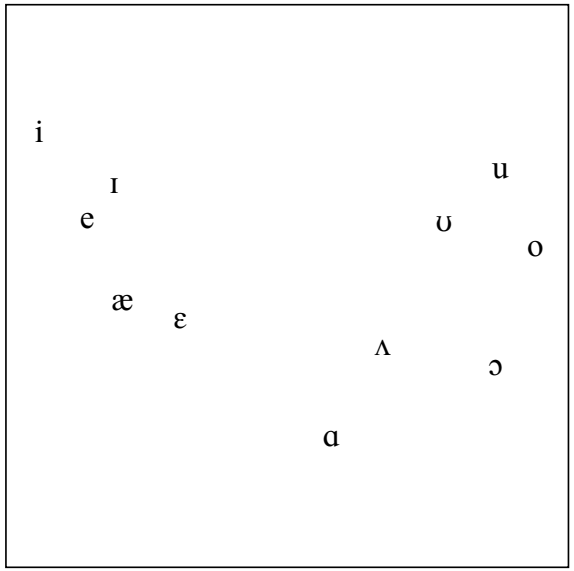
(c) PMI distances

MDS visualization of U.S. English vowels

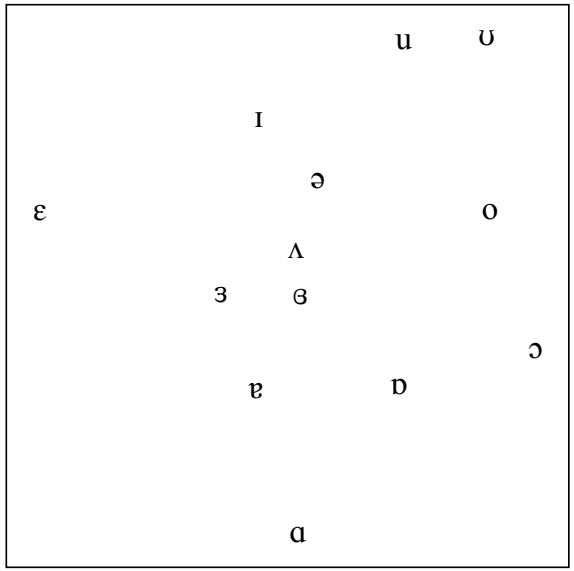
PMI visualization captures 65% of the variation



(a) IPA



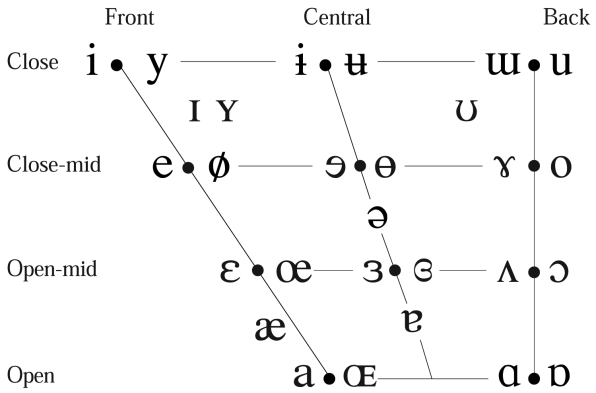
(b) Acoustics



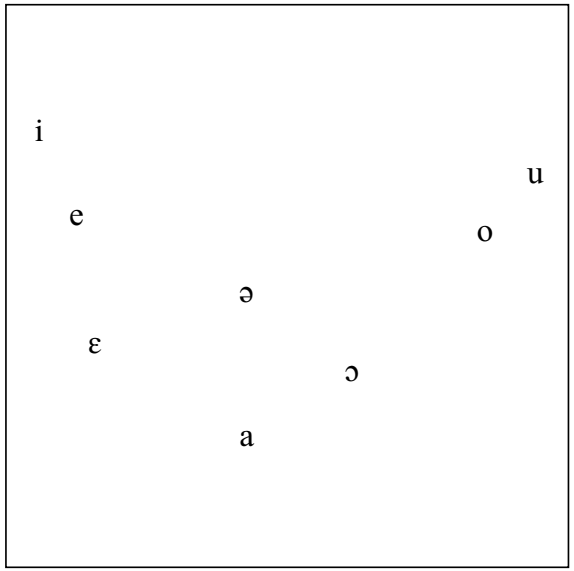
(c) PMI distances

MDS visualization of Bantu vowels

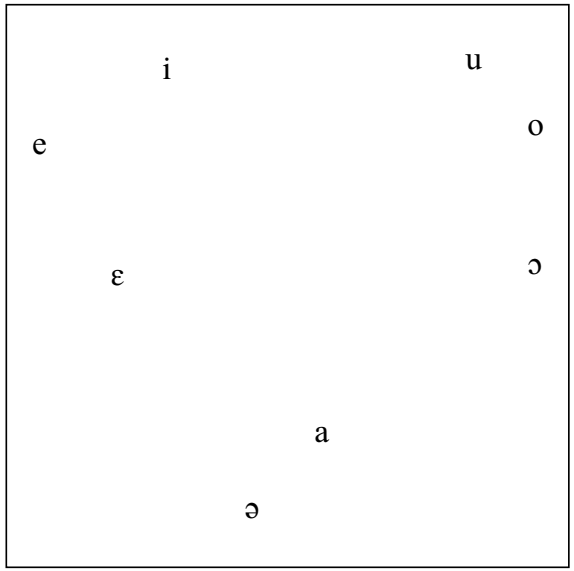
PMI visualization captures 90% of the variation



(a) IPA



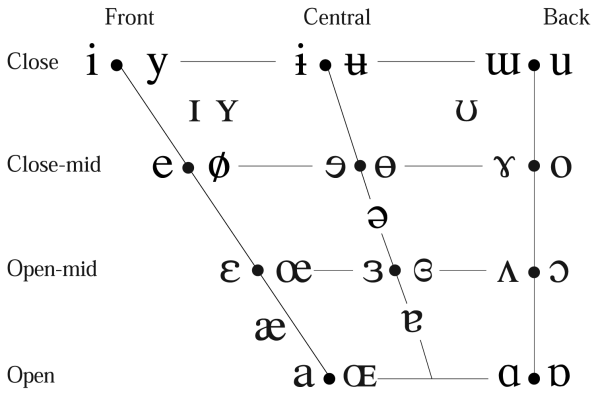
(b) Acoustics



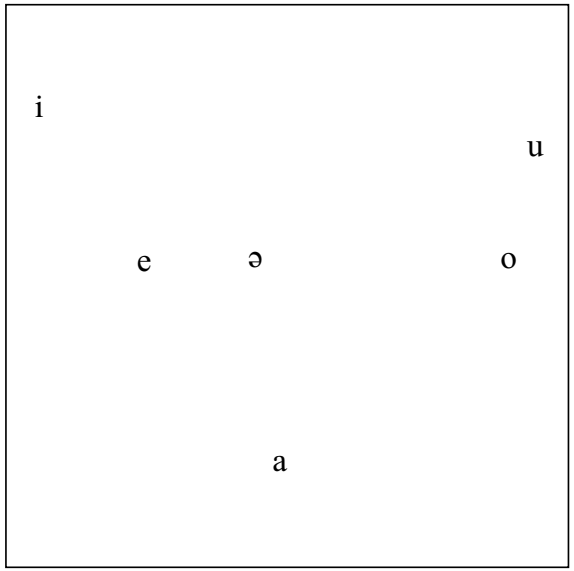
(c) PMI distances

MDS visualization of Bulgarian vowels

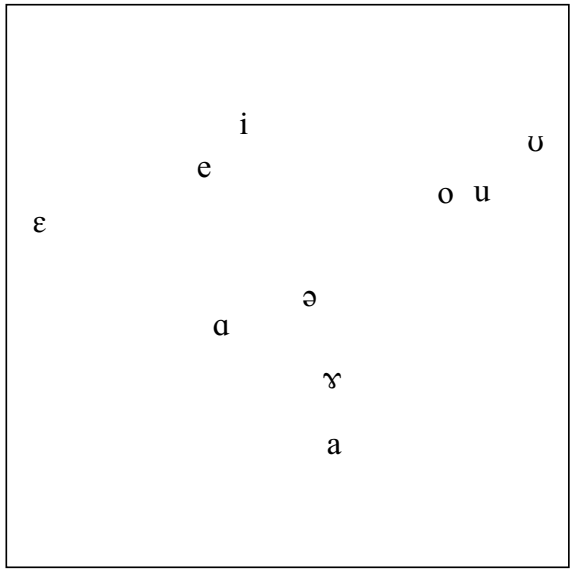
PMI visualization captures 86% of the variation



(a) IPA



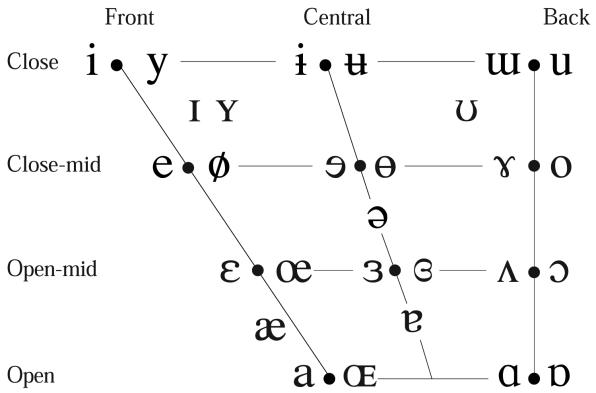
(b) Acoustics



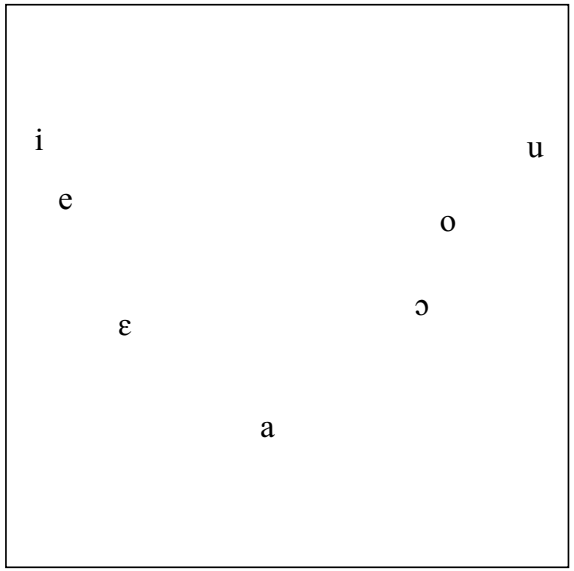
(c) PMI distances

MDS visualization of Tuscan vowels

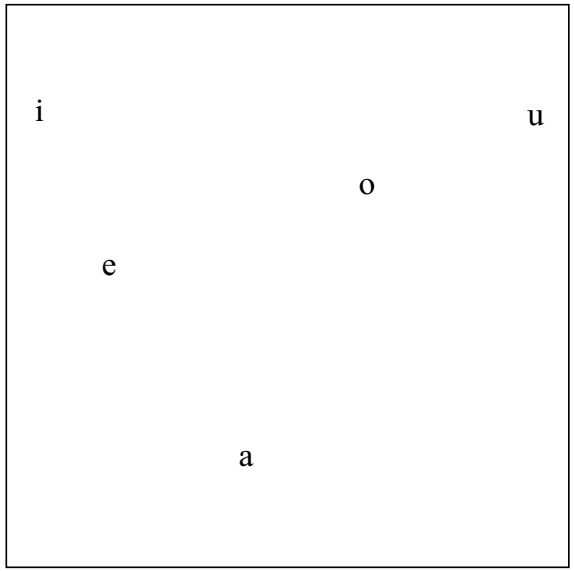
PMI visualization captures 97% of the variation



(a) IPA



(b) Acoustics



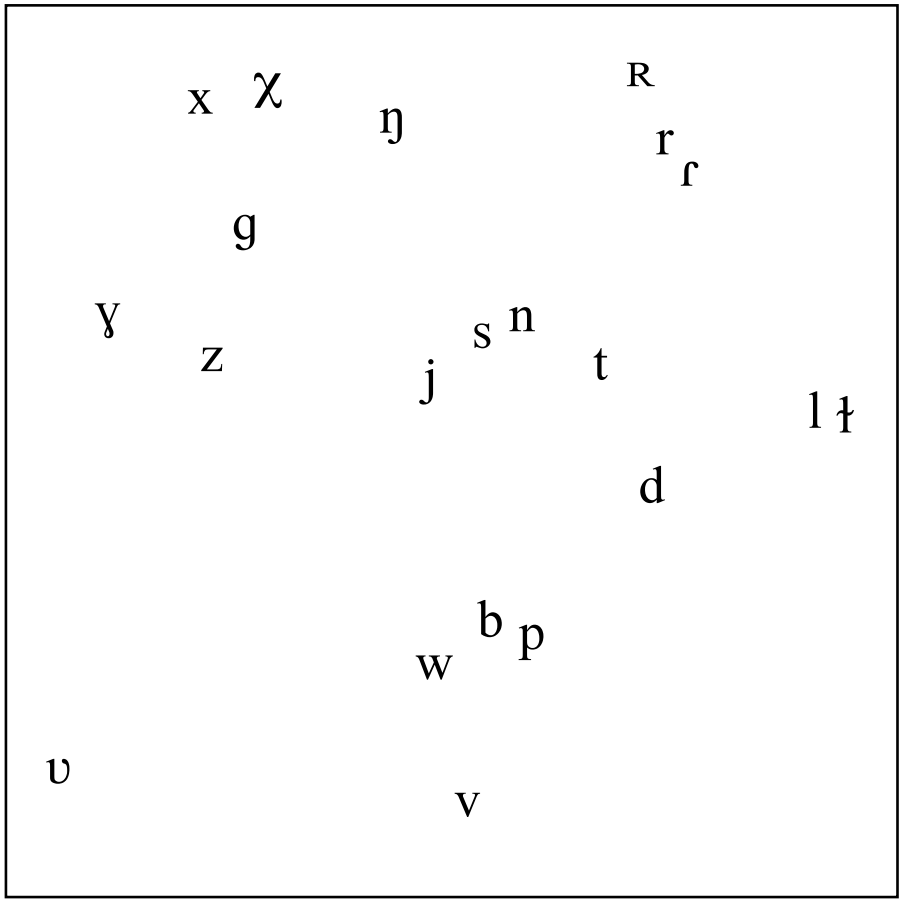
(c) PMI distances

What about consonants?

- Induced distances correlate strongly with acoustic vowel distances
  - Causation is probably the reverse: acoustics explains distributions
    Sweeney's insight: "I gotta use words when I talk to you..."
- But for other segments (consonants) acoustic/phonetic distances are *not* well accepted, and this procedure provides a measure of distance

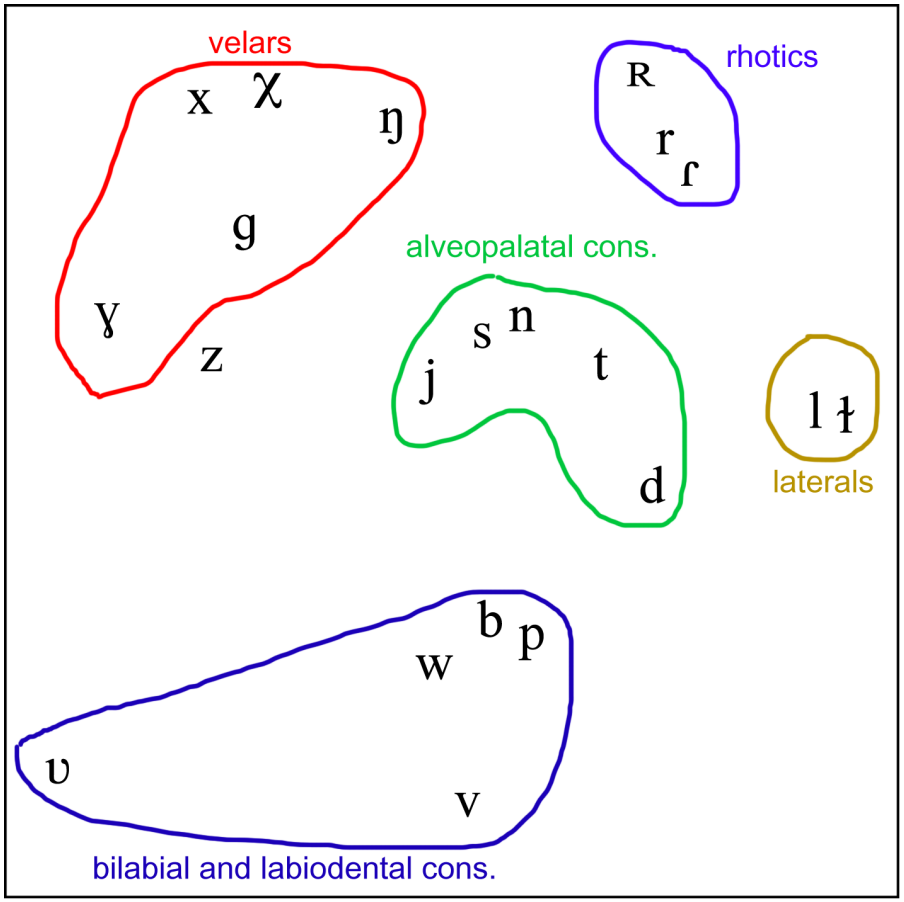
MDS visualization of Dutch consonants

PMI visualization captures 50% of the variation



MDS visualization of Dutch consonants

Place (3 groups) dominates over manner (2 groups) and voicing (no groups)



Conclusion of Part I

- The PMI approach generates sensible sound distances
  - The approach is readily applicable to any (dialect) dataset with similar pronunciations
- For more details and references, see: Martijn Wieling, Eliza Margaretha and John Nerbonne (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2), 307-314.
- In the following, I will apply this method to obtain pronunciation distances on the basis of Dutch dialect data
- Any questions so far?

Conclusion of Part I

- The PMI approach generates sensible sound distances
  - The approach is readily applicable to any (dialect) dataset with similar pronunciations
- For more details and references, see: Martijn Wieling, Eliza Margaretha and John Nerbonne (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2), 307-314.
- In the following, I will apply this method to obtain pronunciation distances on the basis of Dutch dialect data
- Any questions so far?

A sociolinguistic analysis of Dutch dialect distances

- This study attempts to integrate approaches of (social) dialectology and dialectometry in analyzing dialect distances
- Dialectologists mainly focus on social variation but focus on a small number of features (Chambers and Trudgill, 1998)
- Dialectometrists aggregate over many features but mainly focus on dialect geography (e.g., Séguy, 1971; Heeringa and Nerbonne, 2001)
- Here we investigate the effect of geography as well as a number of social factors on dialect distances from standard Dutch for a large set of words in many Dutch dialects
  - We use standard Dutch as our reference variety, as Dutch dialects are known to be converging to the standard language (Van der Wal and Bree, 2008)

Material: pronunciation data

- We used Dutch dialect pronunciations from the GTRP corpus (Goeman and Taeldeman, 1996; Van den Berg, 2003; Wieling et al., 2007)
  - The GTRP is the largest contemporary Dutch dialect data set available
  - It contains transcriptions for 424 locations in the Netherlands
  - The pronunciations were transcribed by several transcribers between 1980 and 1995
  - We used a subset of 559 items having only phonetic variation (mainly verbs, nouns and adjectives; Wieling et al., 2007)
- For all words we obtained:
  - The standard Dutch pronunciation (according to Gussenhoven, 1999)
  - The word frequency (from CELEX; Baayen et al., 1996)

Geographic distribution of locations

Material: social data

- For all locations we obtained:
  - Speaker information (gender and age)
  - Year of recording
  - Average age in the location (in 1995; CBS)
  - Average income in the location (in 1995; CBS)
  - Total number of inhabitants in the location (in 1995; CBS)
  - Male-female ratio in the location (in 1995; CBS)

Methods: determining dialect distances

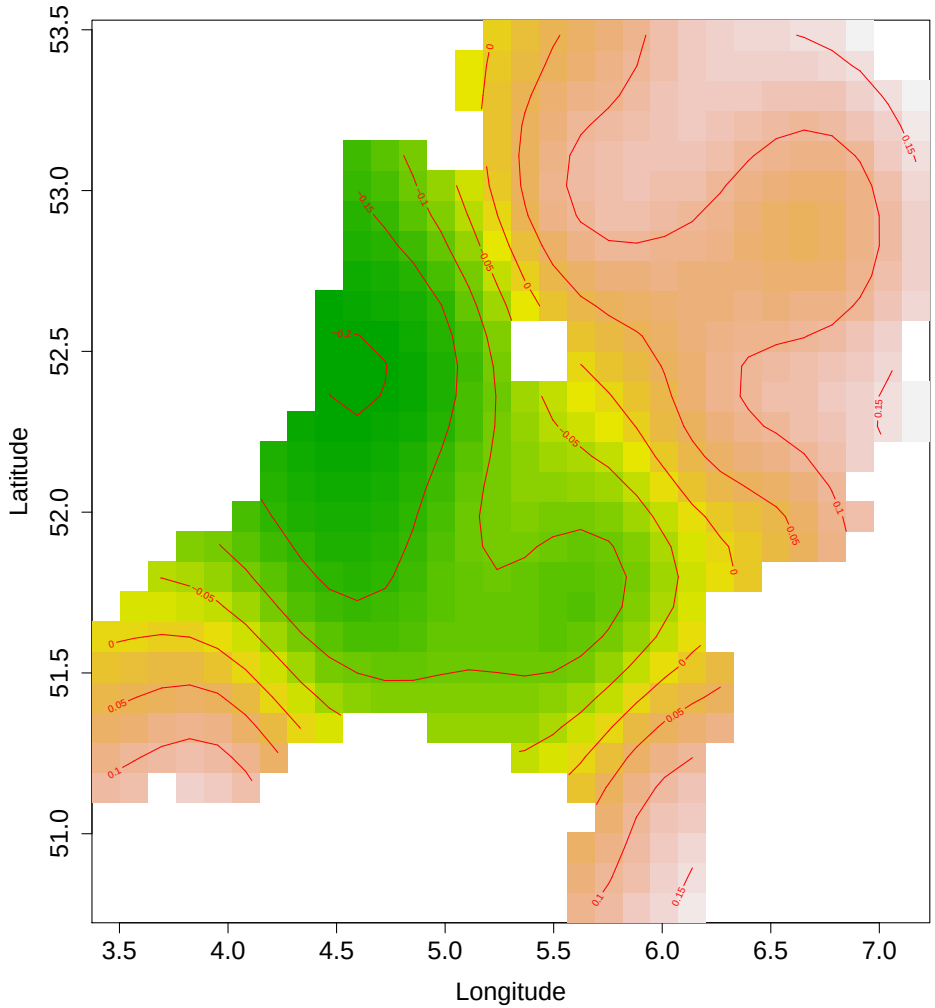
- We use phonetic transcriptions of 562 words in 424 locations in NL
- These are compared to standard Dutch transcriptions using the PMI-based Levenshtein algorithm discussed earlier

m	ɔ		ə		l		k	ə
m			ɛ		l	ə	k	
	0.20		0.15			0.12		0.12

Modeling the influence of geography

- An important determinant for dialect variation is geographic location (people in nearby locations have more contact than in distant locations)
- We include geography by predicting dialect distances with a Generalized Additive Model (GAM) which models the interaction between longitude and latitude
  - The fitted values of this GAM are included as a predictor in our model

Fitted GAM for dialect distance from standard Dutch



Mixed-effects regression extends multiple regression

- Multiple regression: predict one numerical variable on the basis of other independent variables (numerical or categorical)
- We can write a regression formula as $y = I + ax_1 + bx_2 + ...$
- E.g., predict the (centered) linguistic distance from standard Dutch on the basis of word frequency, population size and average population age: $LD = 0 + 0.01WF - 0.005PS + 0.004PA$

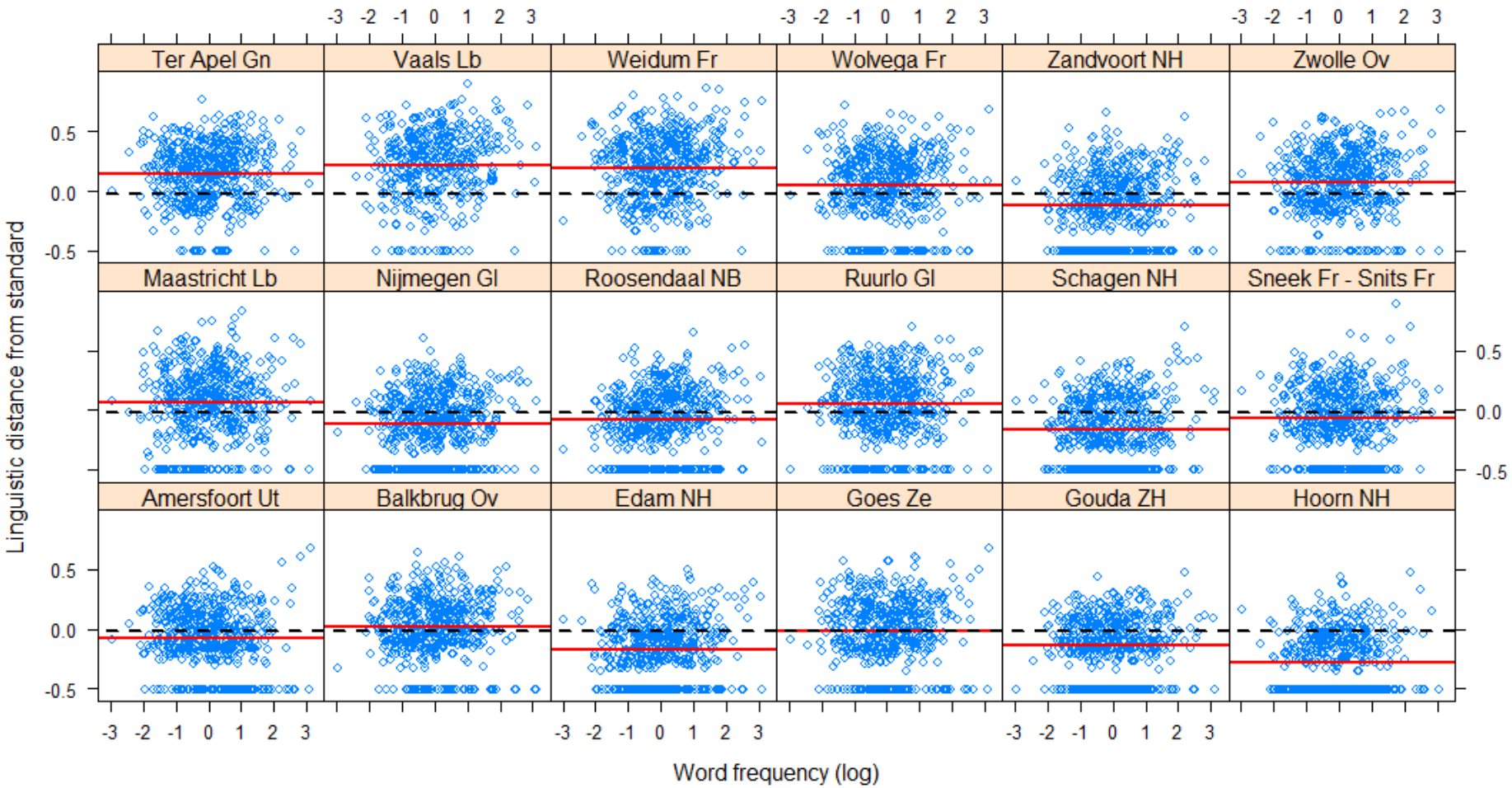
Mixed-effects regression modeling: introduction

- Mixed-effects regression modeling distinguishes fixed effects and random effects
- Fixed effects:
  - Repeatable levels
  - Small number of levels (e.g., gender, word category)
  - Same treatment as in multiple regression (treatment coding)
- Random effects:
  - Levels are a non-repeatable random sample from a larger population
  - Often large number of levels (e.g., location, word, transcriber)

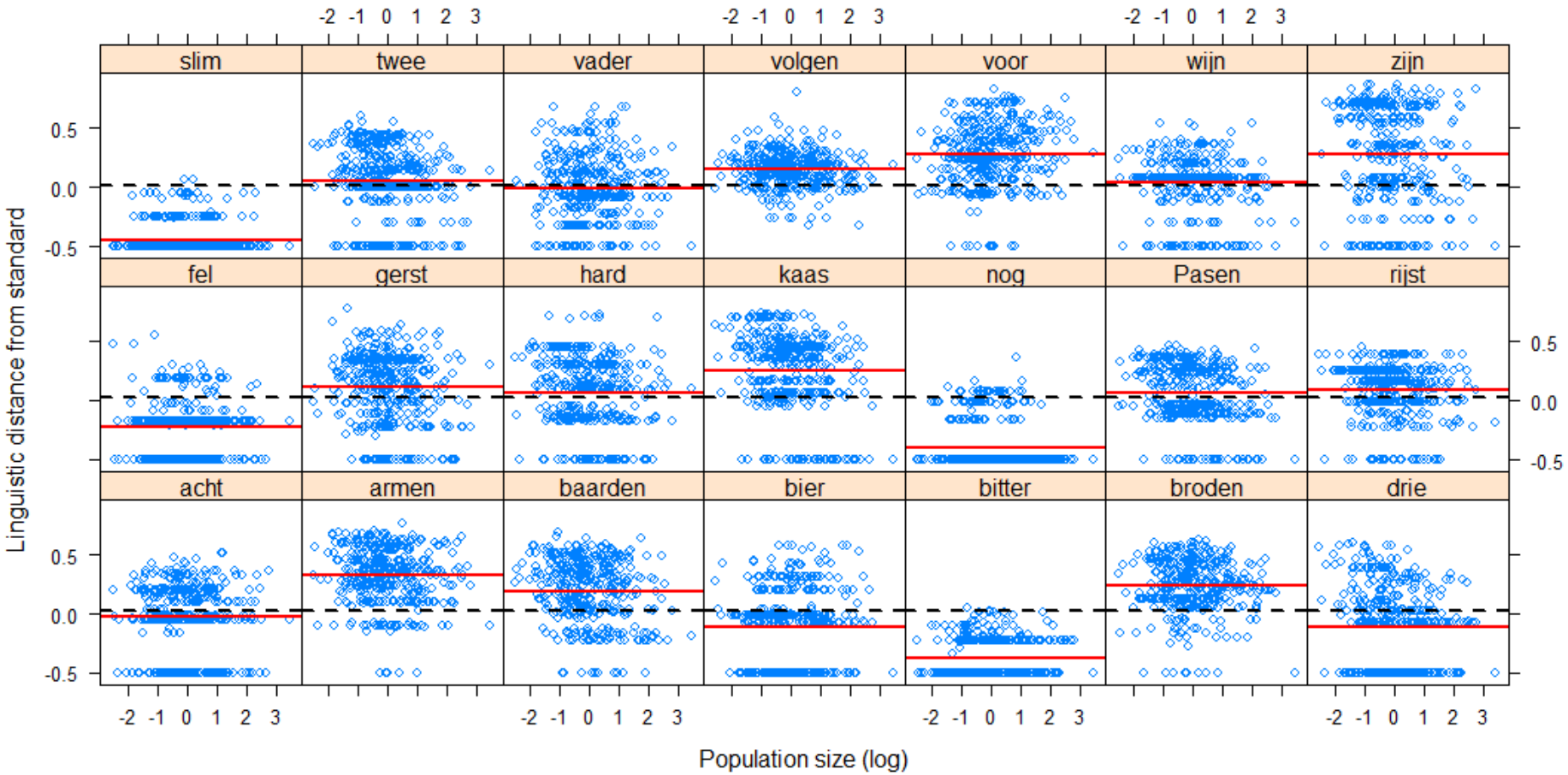
What are random effects?

- Random effects are factors which are likely to introduce systematic variation
  - Some locations have a high linguistic distance (LD) from standard Dutch, while others are close to standard Dutch = Random intercept for location
  - Some words are highly similar to the standard Dutch pronunciation, others are very distinct = Random intercept for word
  - The effect of word frequency on LD might be higher in one location than another (e.g., some dialects may tend to preserve their pronunciation for high frequency words, while others might not)
    = Random slope for word frequency per location
  - The effect of population size on LD might be different for one word than another (e.g., many words might become more similar to standard Dutch in large cities, but some words might show an opposite pattern)
    = Random slope for population size per word

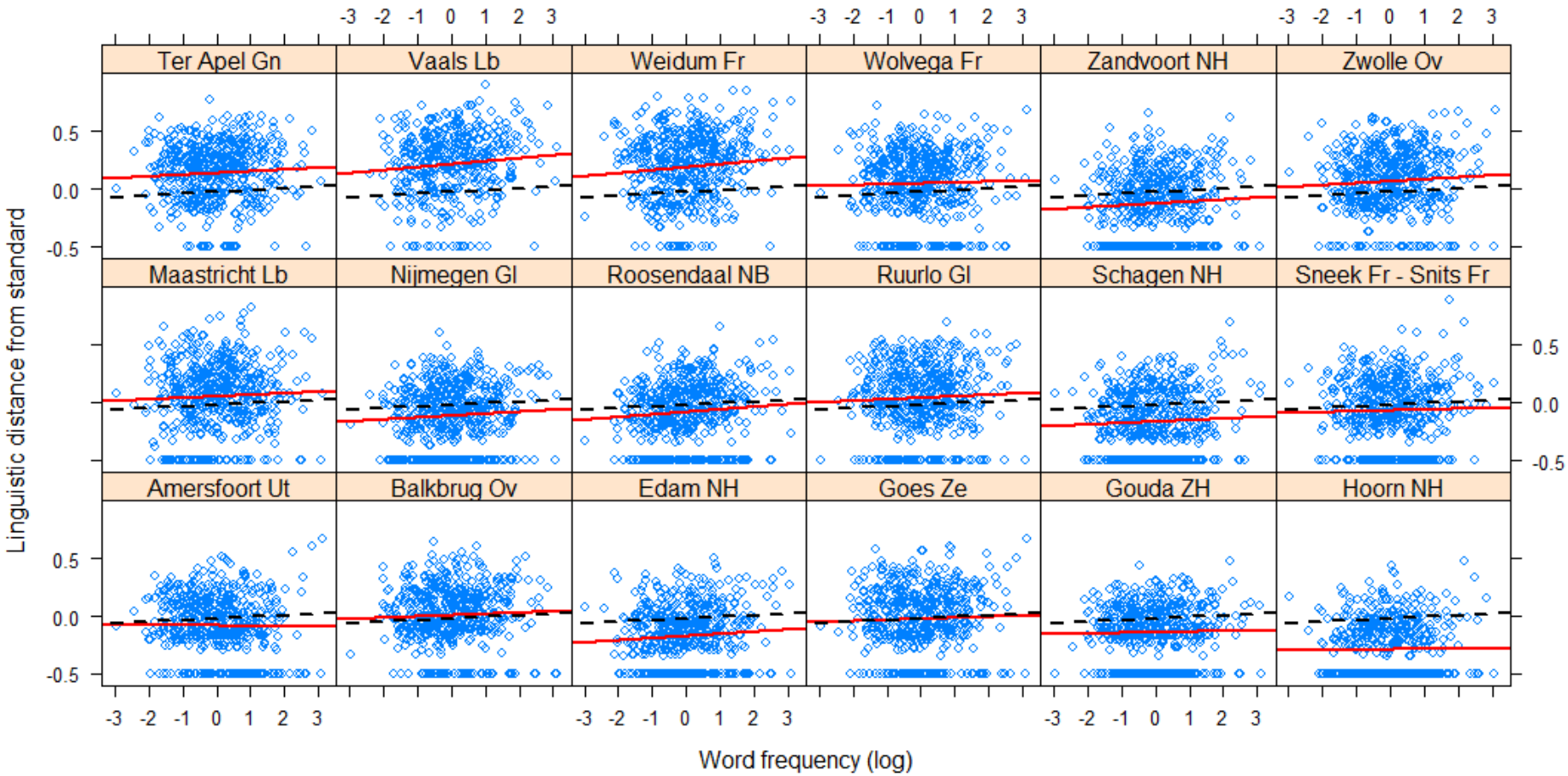
Random intercept for location



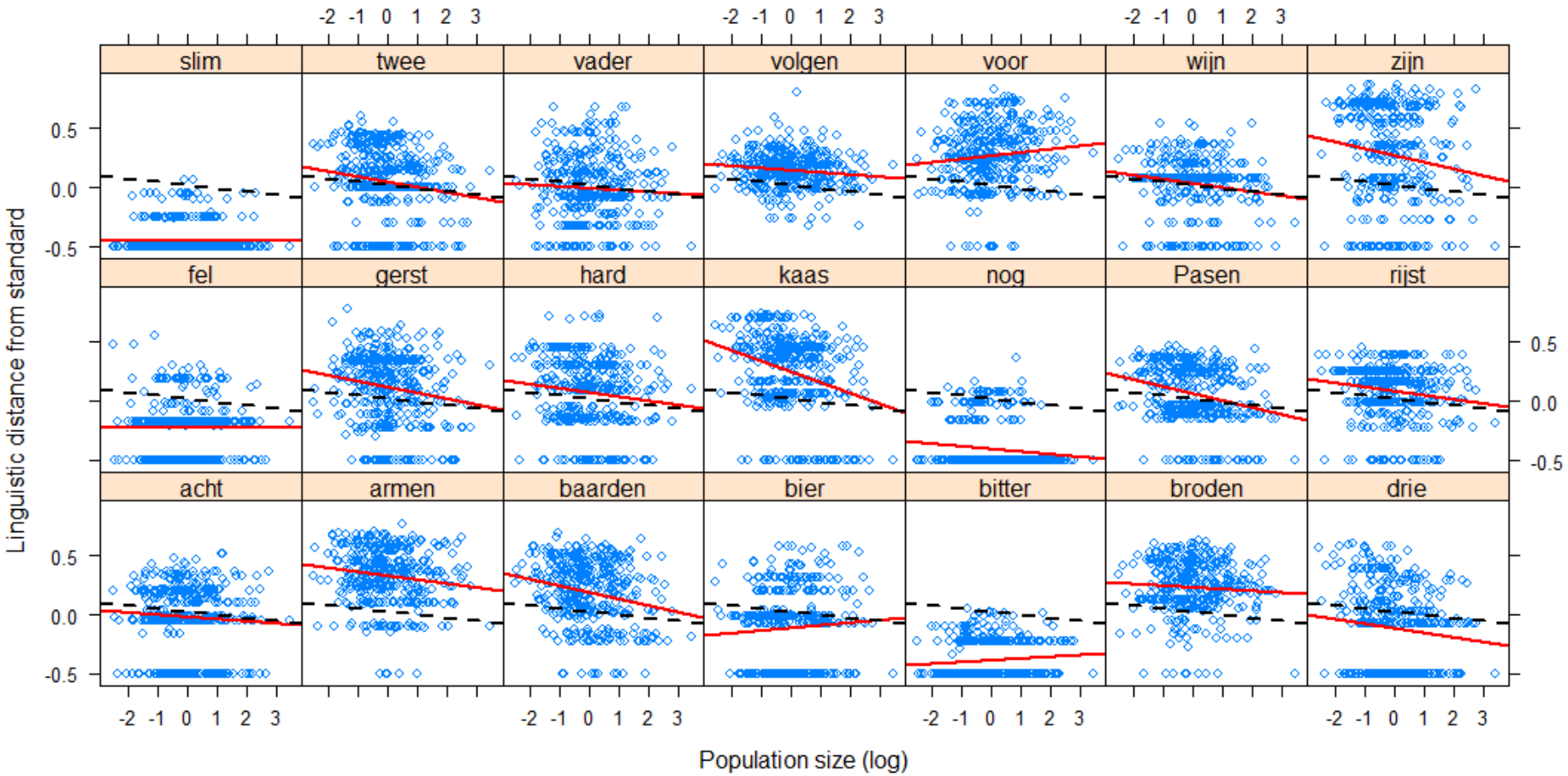
Random intercept for word



Random slope per location



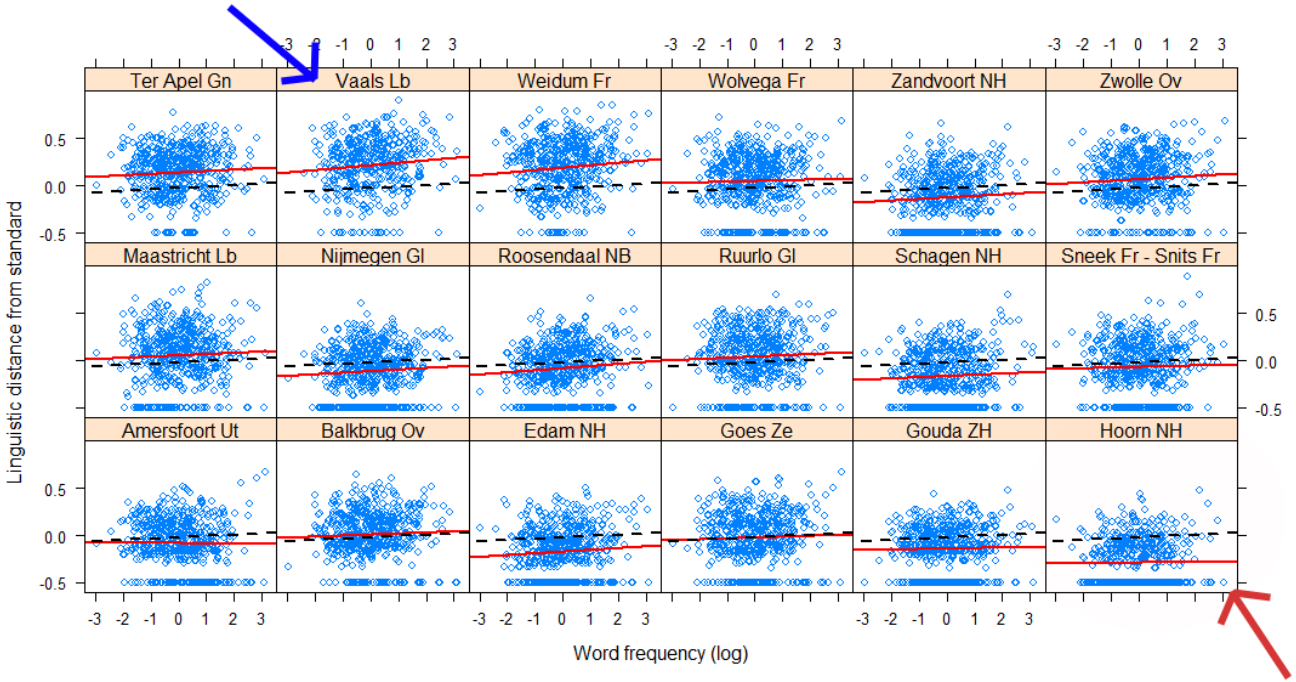
Random slope per word



Specific models for every observation

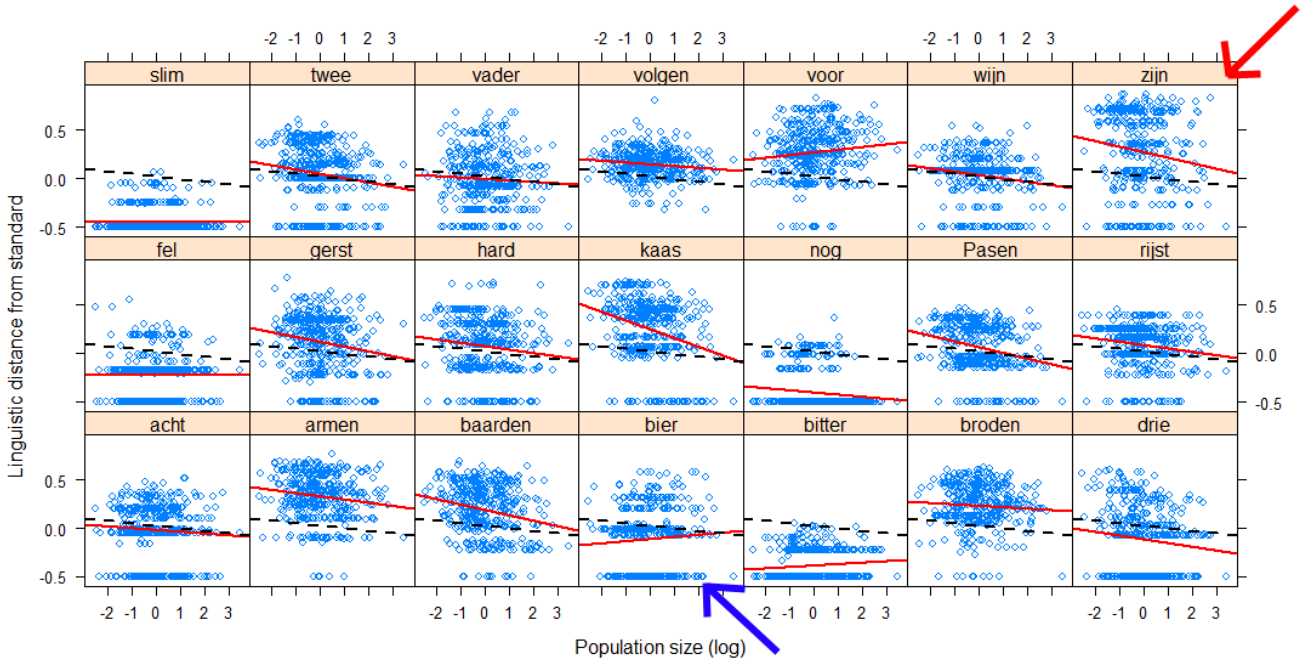
- Mixed-effects regression analysis allow us to use random intercepts and slopes to make the regression formula as precise as possible for every individual observation in our random effects
  - A single parameter (standard deviation) models this variation for every random slope or intercept
  - The actual random intercepts and slopes are derived from this value
  - Likelihood-ratio tests assess whether the inclusion of random intercepts and slopes is warranted

Specific models for every location



- $LD = \quad 0.00 + 0.010\,WF - 0.005\,PS + 0.004\,PA$ (general model)
- $LD = \quad 0.20 + 0.015\,WF - 0.005\,PS + 0.004\,PA$ (Vaals Lb)
- $LD = -0.20 + 0.000\,WF - 0.005\,PS + 0.004\,PA$ (Hoorn NH)

Specific models for every word



- $LD = \quad 0.00 + 0.01\,WF - 0.005\,PS + 0.004\,PA$ (general model)
- $LD = -0.01 + 0.01\,WF + 0.010\,PS + 0.004\,PA$ (word: *bier*)
- $LD = \quad 0.20 + 0.01\,WF - 0.008\,PS + 0.004\,PA$ (word: *zijn*)

Summary of steps

- We investigate which factors predict the dialect distances of 562 words in 424 locations from standard Dutch
- Dialect distances are calculated using the PMI-based Levenshtein distance
- The influence of geography is modeled using a Generalized Additive Model
- We use a mixed-effects regression model
  - Random effects: location, word and transcriber
  - Fixed effects: word frequency, word category, number of inhabitants, average age of inhabitants, ...

Results: fixed effects

	Estimate	Std. Error	*t*-value
Intercept	-0.0153	0.0105	n.s.
GAM distance (geography)	0.9684	0.0274	35.3239
Population size (log)	-0.0069	0.0026	-2.6386
Population average age	0.0045	0.0025	1.8049
Population average income (log)	-0.0005	0.0026	n.s.
Noun instead of Verb/Adjective	0.0409	0.0122	3.3437
Word frequency (log)	0.0198	0.0060	3.2838
Vowel-consonant ratio (log)	0.0625	0.0059	10.5415

*t-values indicate significance ($p < 0.05$) if $|t| > 2$ (two-tailed) or $|t| > 1.65$ (one-tailed)

Significant demographic predictors

- Geography
  - Nearby varieties tend to speak more similar dialects (Nerbonne and Kleiweg, 2007)
- Population size: larger cities (irrespective of geographical location) have a pronunciation closer to standard Dutch
  - People have weaker ties in urban populations, causing dialect leveling (Milroy, 2002)
- Average population age: Locations with a younger population have a pronunciation closer to standard Dutch
  - Younger people speak less dialectal than older people (Van der Wal and Bree, 2008)
- The effect of these predictors varies per word

Significant lexical predictors

- Vowel-to-consonant ratio: words with relatively many vowels have a pronunciation distant from standard Dutch
  - Vowels are much more variable than consonants (Keating et al., 1994)
- Word frequency: more frequent words are more distant from standard Dutch
  - More frequent words are more resistant to change (Pagel et al., 2007)
- Word category: nouns are more distant from standard Dutch than verbs and adjectives
  - Nouns are more resistant to change than verbs and adjetives (Pagel et al., 2007)
- The effect of these predictors varies per location

Results: random effects

Factors	Rnd. effects	Std. Dev.	Cor.	
Word	Intercept	0.1394		
	Pop. size (log)	0.0186		
	Pop. avg. age	0.0086	-0.856	
	Pop. avg. income (log)	0.0161	0.867	-0.749
Location	Intercept	0.0613		
	Word freq. (log)	0.0161	-0.084	
	Noun instead of Verb/Adjective	0.0528	-0.595	0.550
Transcriber	Intercept	0.0260		
Residual		0.2233		

*The inclusion of all random slopes and intercepts was warranted by likelihood-ratio tests

*A richer random effect structure is likely possible, but not computationally feasible (now: 24 CPU hours!)

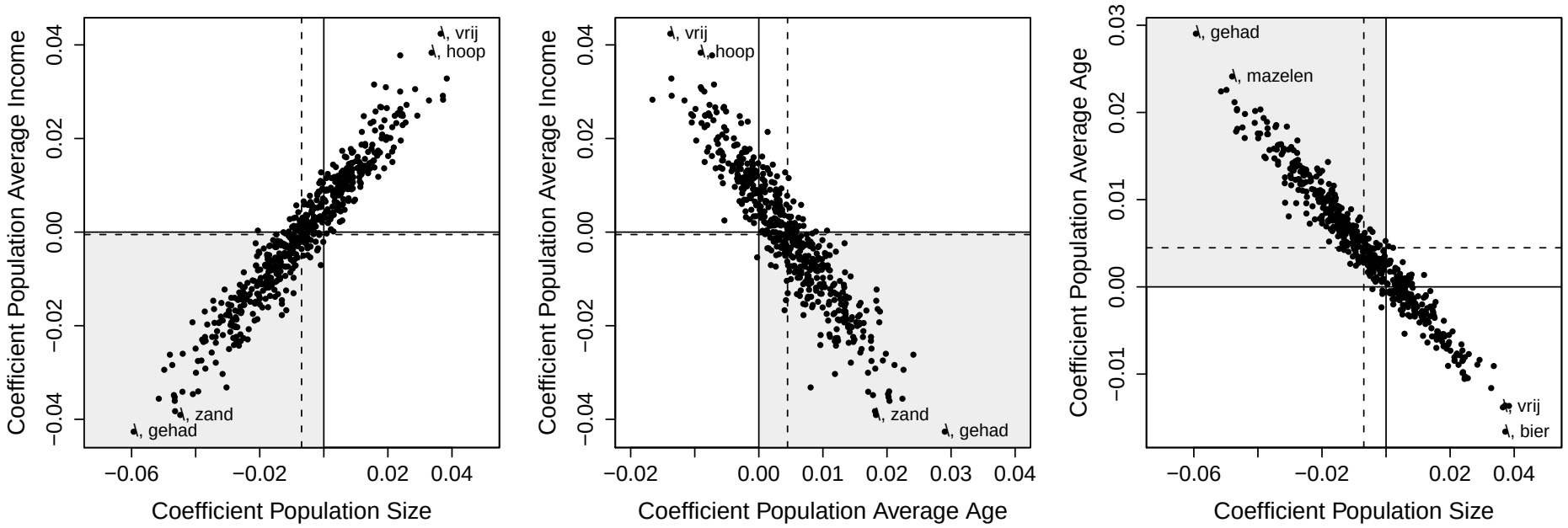
Final model: by-word random slopes correlate highly

Factors	Rnd. effects	Std. Dev.	Cor.	
Word	Intercept	0.1394		
	Pop. size (log)	0.0186		
	Pop. avg. age	0.0086	-0.856	
	Pop. avg. income (log)	0.0161	0.867	-0.749
Location	Intercept	0.0613		
	Word freq. (log)	0.0161	-0.084	
	Noun instead of Verb/Adjective	0.0528	-0.595	0.550
Transcriber	Intercept	0.0260		
Residual		0.2233		

*The inclusion of all random slopes and intercepts was warranted by likelihood-ratio tests

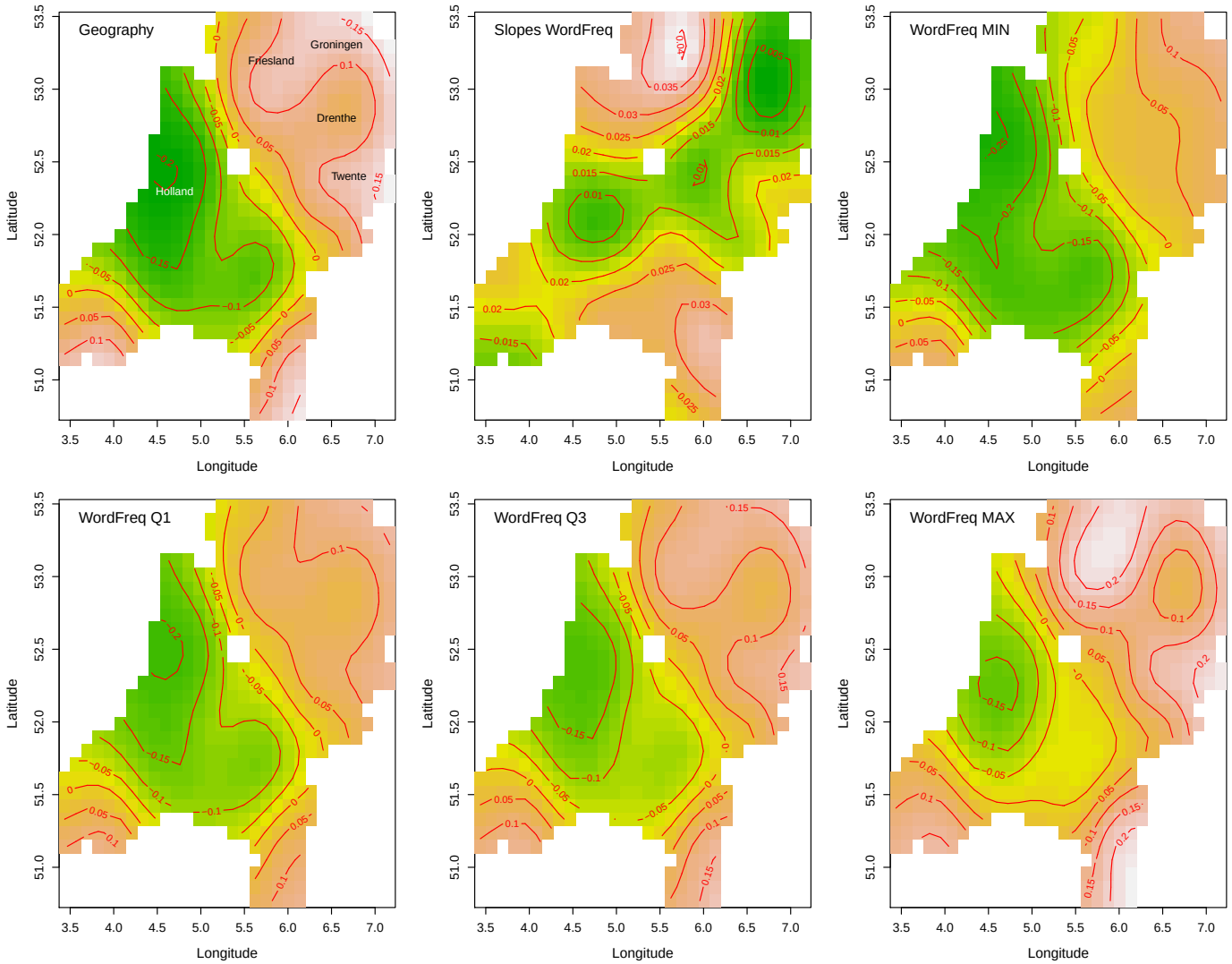
*A richer random effect structure is likely possible, but not computationally feasible (now: 24 CPU hours!)

Correlation structure of by-word random slopes

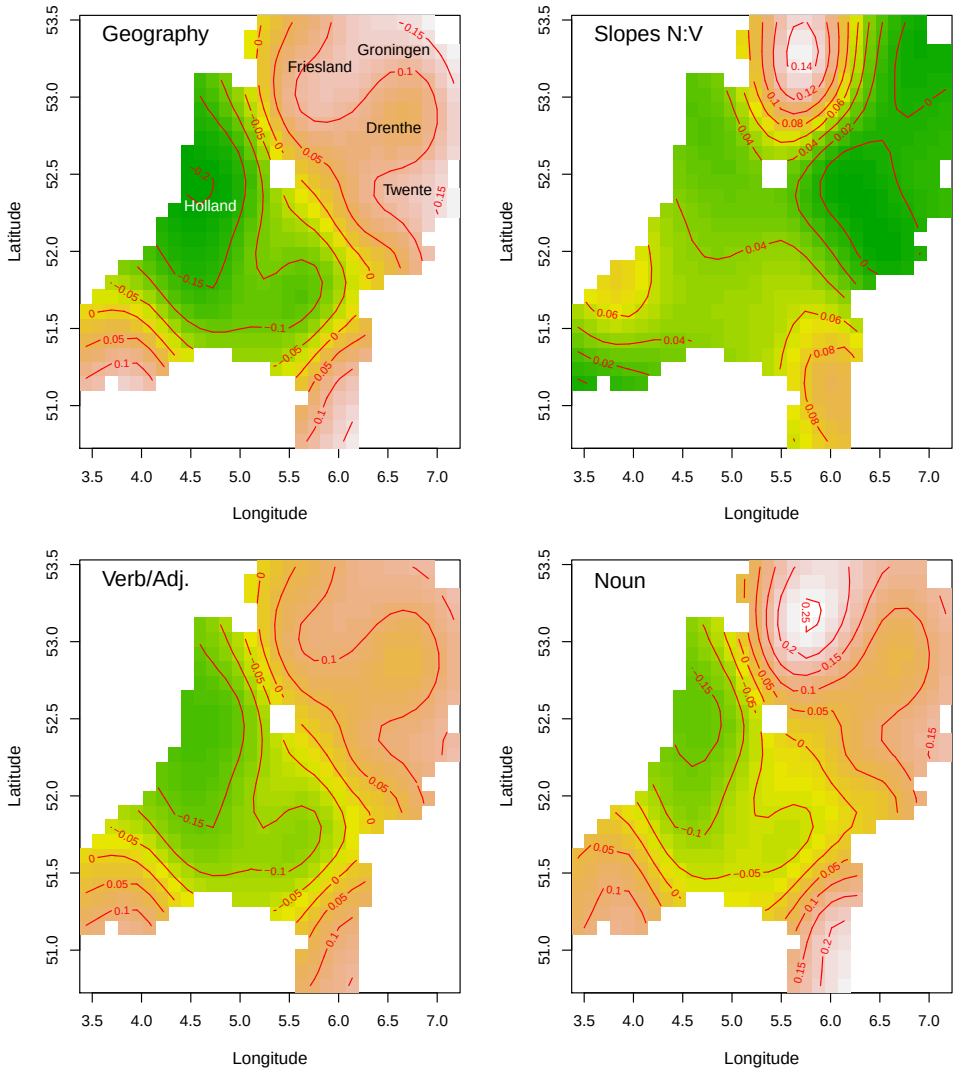


- $LD = -0.0069PS - 0.0005PI + 0.0045PA + \ldots$ (general model)
- $LD = -0.0600PS - 0.0420PI + 0.0290PA + \ldots$ (*gehad*: extreme pattern)
- $LD = \ \ 0.0380PS + 0.0420PI - 0.0110PA + \ldots$ (*vrij*: inverted pattern)

By-location random slopes for word frequency



By-location random slopes for the Noun-Verb contrast



Concluding remarks about random intercepts and slopes

- Words and locations show much variation
  - Random intercepts for word, location and transcriber are necessary
- The effect of several word-related variables differs per location
  - By-location random slopes are necessary
- The effect of several demographic variables differs per word
  - By-word random slopes are necessary
- Including these random effects is essential to ensure our fixed effects are valid

Discussion

- Our model “explained” about 45% of the variation in the data with respect to the distance from standard Dutch
- We identified a number of location- and word-related variables playing an important role in predicting the dialect distance from standard Dutch
  - Geography (i.e. social contact between locations)
  - Location-related factors: population size and average age
  - Word-related factors: word category, word frequency and vowel-cons. ratio
- Using a mixed-effects regression approach ensures our results are generalizable and enabled us to quantify and study the variation of individual words and speakers

Conclusion of Part II

- We illustrated a promising approach combining the merits of dialectology (investigating social factors) and dialectometry (using a large set of items, and including geography) to investigate dialect variation at the word pronunciation level
  - The method is not only applicable to pronunciation data, but also to lexical data using logistic regression (Wieling, Montemagni, Nerbonne and Baayen, submitted)

- For more details and references, see: Martijn Wieling, John Nerbonne and R. Harald Baayen (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLoS ONE*, 6(9): e23613. doi:10.1371/journal.pone.0023613.

Thank you for your attention!