University at Buffalo *The State University of New York* | REACHING OTHERS

Autologistic Regression in Linguistic Typology

Yoshihiko Asao
yoshihik@buffalo.edu

Introduction

- A typological frequency difference is often taken as a linguistic preference and given linguistic explanations
- However, there are often large-scale geographical patterns
- It is difficult to distinguish a true linguistic preference from a historical accident

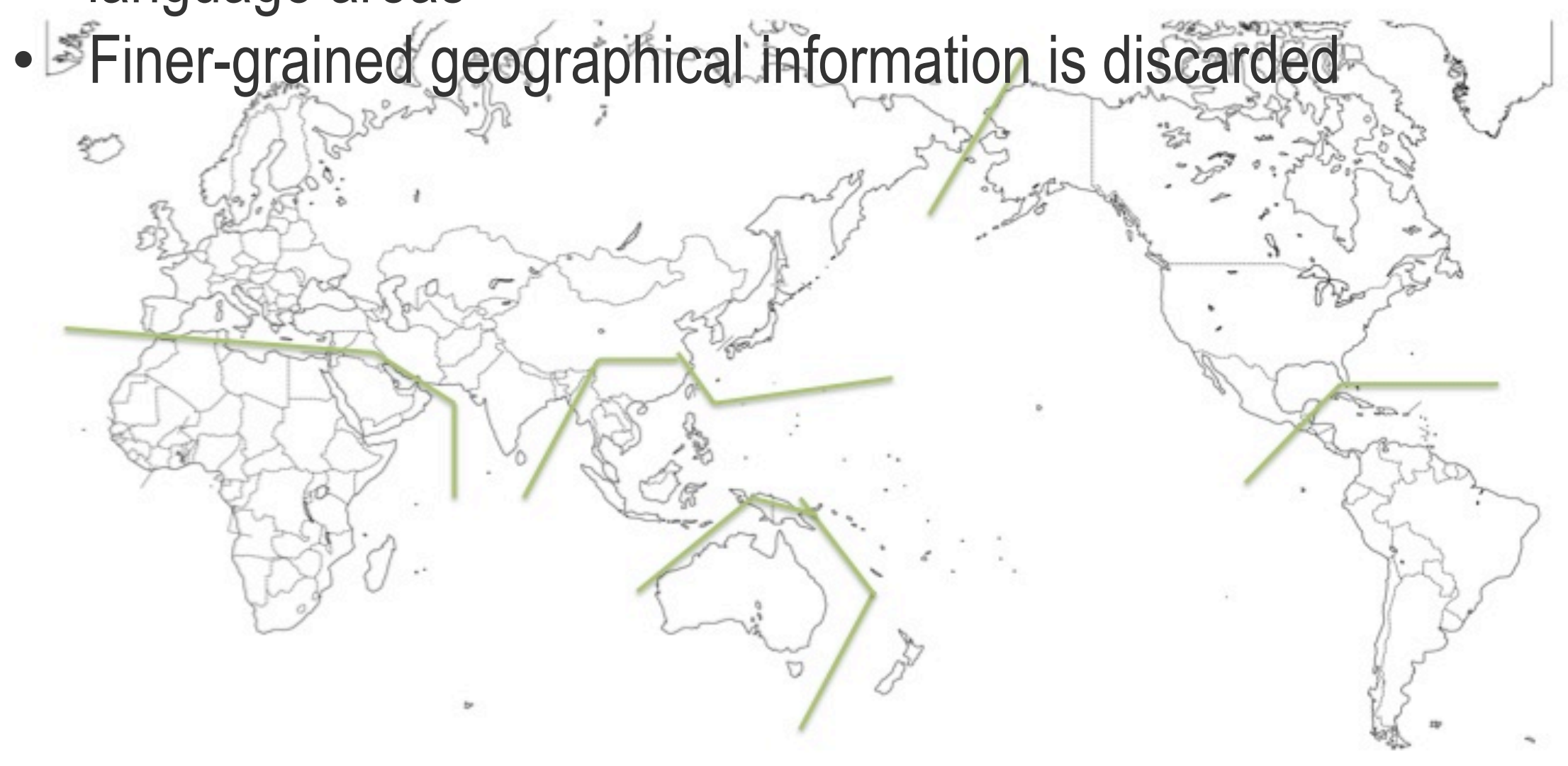
Previous Approaches

Independent sample approach (Perkins 1989, among others)

- Needs to discard most of the data

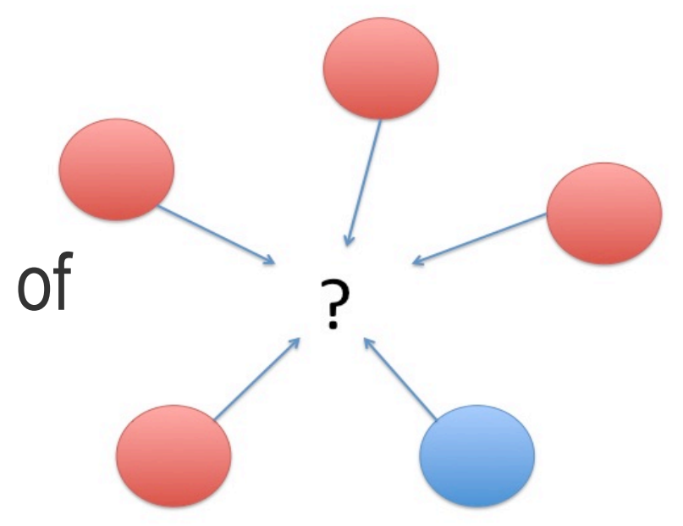
Language area approach (Dryer 1989, 1992, Bickel 2008)

- Arbitrariness and potential interdependence between language areas
- Finer-grained geographical information is discarded



Autologistic Regression

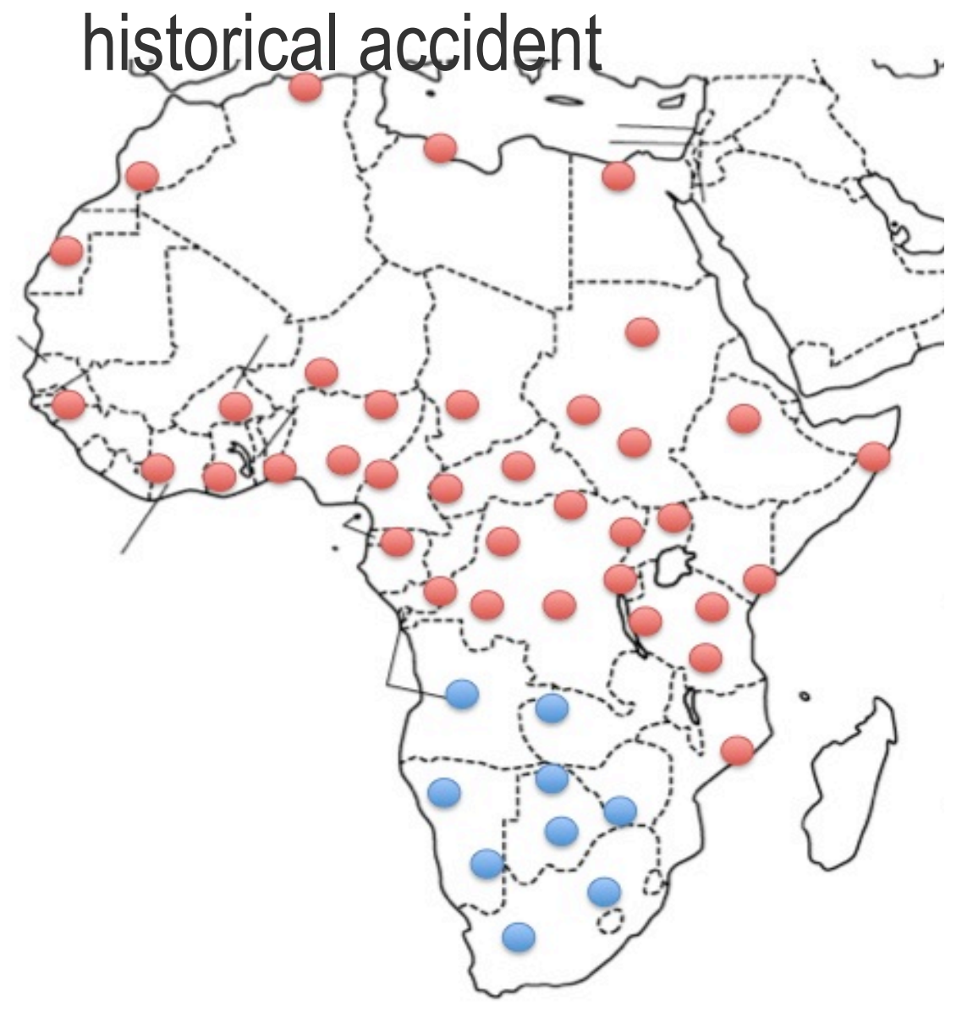
- Similar to the logistic regression in Bickel (2008)
- Instead of language areas, the opinions from neighbors are a part of the model
- Inspired by discussions on similar issues in ecology (Dormann 2007)



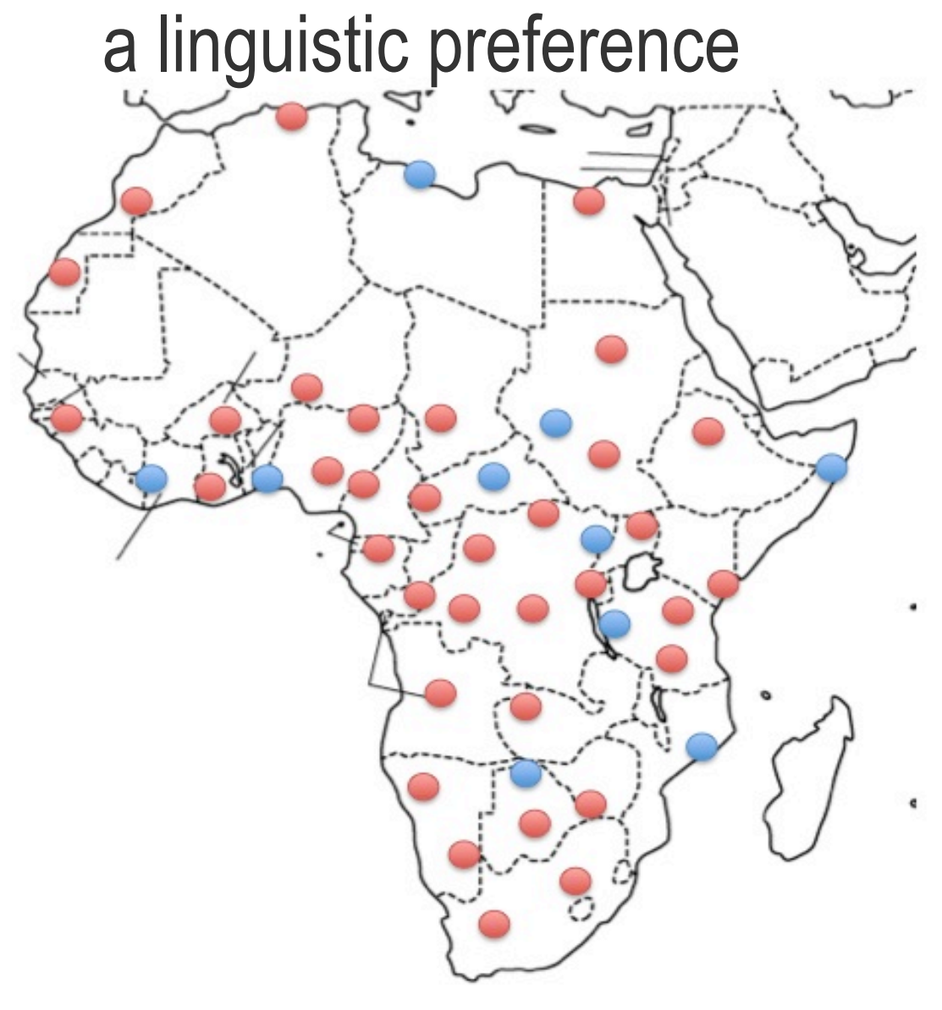
Autologistic Regression (cont.)

- Key idea: if most variance is explained away as the retention of the features of related languages, the evidence for the universal linguistic preference is weak.

More predictable from neighbors: The pattern can be explained by a few historical accident



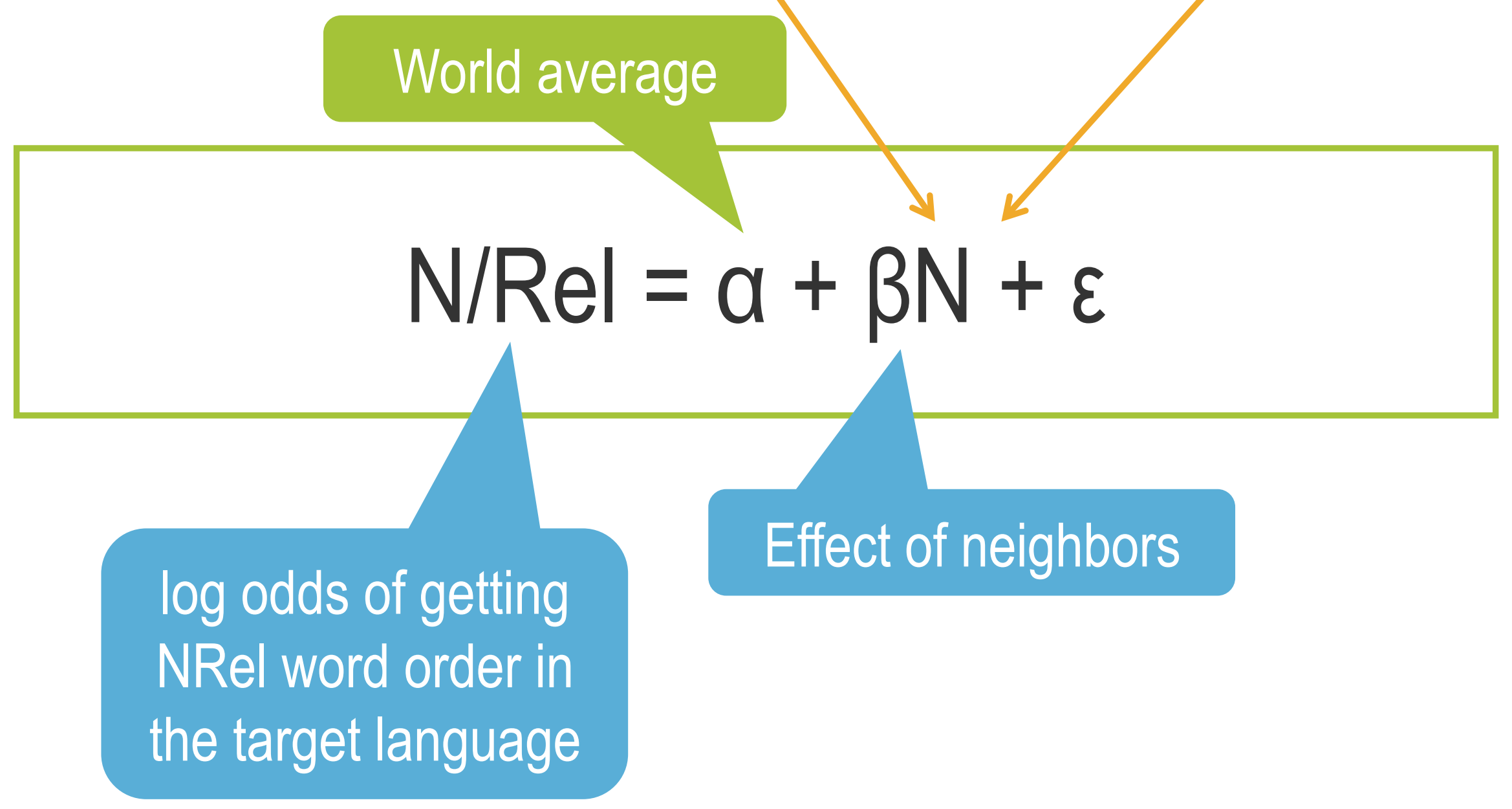
Less predictable from neighbors: Spontaneous change is frequent, suggesting the frequency difference is more likely to be a linguistic preference



Example (Order of relative clause and noun)

English	
Five closest languages	
Welsh	NRel
Romani (Welsh)	NRel
Frisian	NRel
Cornish	NRel
Dutch	NRel
# of NRel	5
z-score	0.593

Japanese	
Five closest languages	
Ainu	RelN
Korean	RelN
Dagur	RelN
Nivkh	RelN
Seediq	NRel
# of NRel	1
z-score	-1.840



World average

$$N/Rel = \alpha + \beta N + \varepsilon$$

log odds of getting NRel word order in the target language

Effect of neighbors

Procedure

- Data values and geographical distances are taken from WALS chapters (Dryer and Haspelmath 2011)
- Find the best model using stepwise regressions with AIC

Results

Examples from Phonology
CLICK

	AIC	pR^2
Click ~ I	90.6	
★ Click ~ Neighbor	28.7	72.1%

AIC = Akaike Information Criteria
pR^2 = McFadden's pseudo-R squared

TH-SOUND (non-sibilant dental or alveolar fricative)

	AIC	pR^2
Th-sound ~ I	306.5	
★ Th-sound ~ Neighbor	303.5	1.6%

- Although both click sounds and th-sounds are typologically rare features, the former is much more predictable from neighboring languages
- The rarity of th-sound is more likely to reflect a universal preference

Example from Syntax
N/Rel (Order of relative clause and noun)

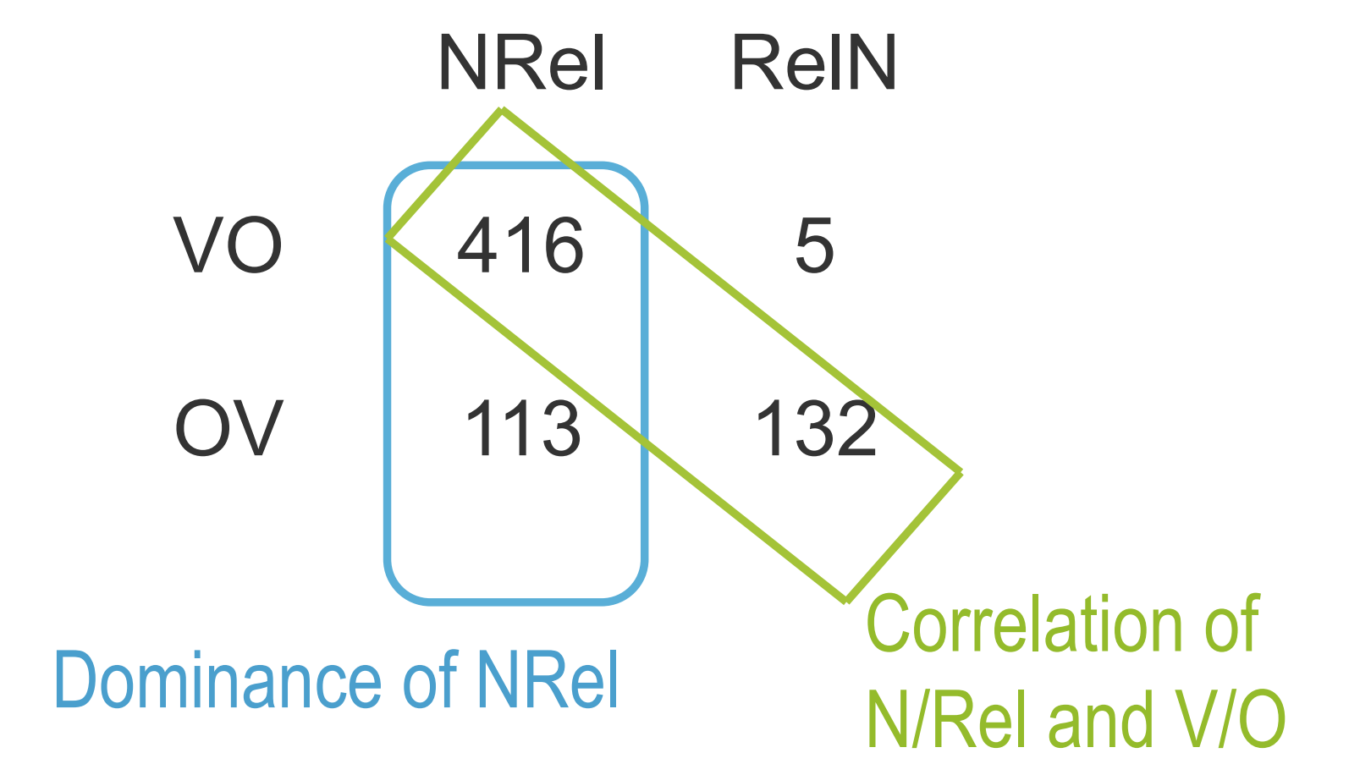
	AIC	pR^2
N/Rel ~ I	714.6	
★ N/Rel ~ Neighbor	325.6	54.3%

Discussion

- Lack of random sampling: the use of parametric statistics may not be appropriate
- The model lacks the distinction between geographical and genealogical factors
- Autologistic regression is not without criticism (Dormann 2007)

Results: Implicational Universals

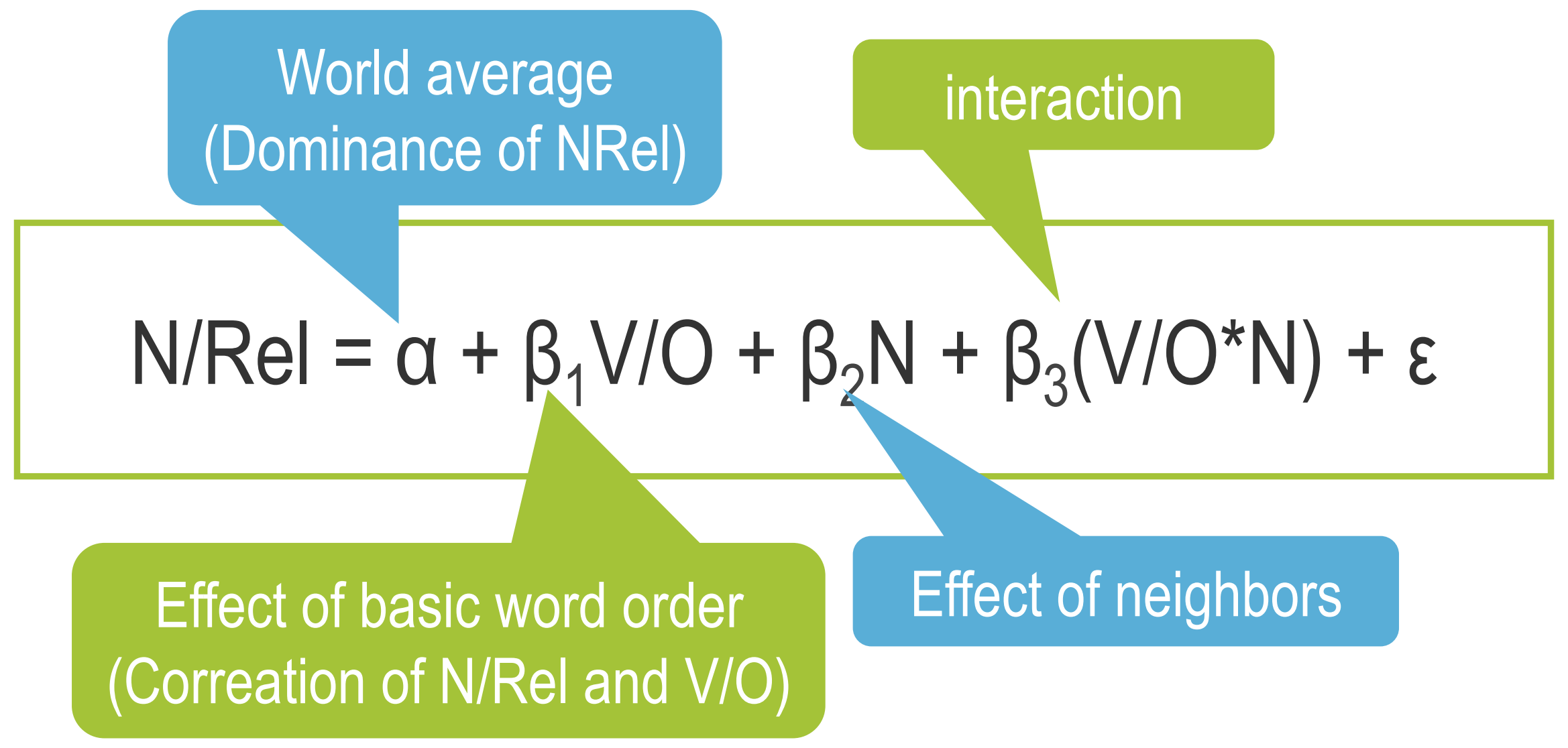
NRel if VO



	NRel	RelN
VO	416	5
OV	113	132

Dominance of NRel

Correlation of N/Rel and V/O



World average (Dominance of NRel)

interaction

$$N/Rel = \alpha + \beta_1 V/O + \beta_2 N + \beta_3 (V/O*N) + \varepsilon$$

Effect of basic word order (Correation of N/Rel and V/O)

Effect of neighbors

	AIC	pR^2
N/Rel ~ I	678.9	
N/Rel ~ Neighbor	322.7	52.9%
N/Rel ~ VO	396.4	42.0%
★ N/Rel ~ Neighbor + VO	254.9	63.2%
N/Rel ~ Neighbor + VO + neighbor*VO	256.4	63.3%

Conclusion

Autologistic regression may be a useful method to discern a true linguistic preference from a historical accident

References

Bickel (2008) A general method for the statistical evaluation of typological distributions. Draft.

Dormann (2007) Assessing the validity of autologistic regression. *Ecological Modelling* 207, 234-242.

Dryer (1989) Large linguistic areas and language sampling. *Studies in Language* 13: 2, 257-292.

Dryer (1992) Greenbergian word order correlations. *Language* 68: 1, 81-138.

Dryer and Haspelmath (2011) *World Atlas of Language Structures Online*.

Perkins (1989) Statistical techniques for determining language sample size. *Studies in Language* 13: 2, 293-315.

Linguistic Society of America (Jan 2013) Minneapolis, MN