Autologistic Regression in Linguistic Typology

Yoshihiko Asao*

LSA Annual Meeting, Minneapolis, January 2–5, 2014

1. Background. Typological frequency biases are often taken as linguistic preferences, and prompt research on why such preferences exist, either formal or functional. However, it is often difficult to tell whether a given asymmetry is truly a reflection of a linguistic preference, or just a result of historical accidents (Dryer, 1989). For example, the fact that more languages postpose a relative clause has been explained by Hawkins' (1994; 2004) processing-based theory, among others. However, languages with preposed relative clauses are disproportionally concentrated in Eurasia (Dryer, 2011). This may indicate that the current typological frequencies are a mere result of historical accidents, not a linguistic preference.

A number of proposals have been made to discern true linguistic preferences from historical accidents. One approach is to create a sample of distant languages so that no pair of languages is historically related. Such an approach, however, needs to throw away most of the data points that could otherwise be useful. More seriously, there is no guarantee that a sample of independent languages can be constructed; all languages in the world might be under the influence of a single historical event (Maslova, 2000).

Another approach is to divide the world into pre-defined linguistic areas, and see whether there is a frequency bias independent of areas (Dryer, 1989, 1992; Bickel, 2008). Dryer's approach uses a non-parametric test, while Bickel includes linguistic areas as an explanatory variable in logistic regression. These approaches also suffer from the loss of information, because they are only sensitive to the geographical scale we choose. For example, languages with phonemic clicks concentrate in Southern Africa. This may indicate that their typological frequency is a historical accident, as opposed to cases where languages with phonemic clicks scatter around the continent. However, the model cannot be sensitive to this fact when we treat Africa as a single linguistic area and discard finer geographic information.

2. Autologistic regression. To overcome these issues, this paper applies autologistic regression analysis to linguistic typology, a common method to model geographically correlated data in geography and ecology (Dormann, 2007). In autologistic regression, neighbors' responses are used to predict the response at issue. For example, whether a language preposes a relative clause will be predicted by examining whether the surrounding languages prepose a relative clause. When the responses of too many languages can be predicted in this way, it means that the worldwide distribution can be readily explained by the retention of the feature of closely related languages, and the evidence for the linguistic preference will be weakened.

Author: Yoshihiko Asao, University at Buffalo (asaokitan@gmail.com)

In this study, five geographically closest languages are used to predict the response, based on the coordinate data from the World Atlas of Language Structures (Dryer & Haspelmath, 2011). Suppose, for example, that we would like to know whether the head noun precedes or follows a relative clause (N/Rel) in a given language. Let N be the neighbors' average response normalized as a z-score. Then the model can be represented as follows.

$$\text{N/Rel} = \alpha + \beta N + \epsilon$$

Here α will be the world average, β the coefficient that represents how strong the neighbor factor is, and ϵ the error term.

For example, the five languages geographically closest to English are Welsh, Romani, Frisian, Cornish and Dutch. All of these five languages postpose relative clauses (NRel), which we regard as the raw score of 5. By normalizing it we will obtain the z-score of 0.593. On the other hand, the five closest languages of Japanese are Ainu, Korean, Dagur, Nivkh and Seediq, all of which are RelN except Seediq. This procedure is summarized in Table 1.

Table 1: Calculation of the neighbor factor

	English		Japanese	
Five closest languages	Welsh	NRel	Ainu	RelN
	Romani (Welsh)	NRel	Korean	RelN
	Frisian	NRel	Dagur	RelN
	Cornish	NRel	Nivkh	RelN
	Dutch	NRel	Seediq	NRel
raw score	5		1	
z-score	0.593		-1.840	

Table 2 compares the results of the autologistic regressions for clicks and th-sounds (non-sibilant dental and alveolar fricatives) based on the data of 567 languages in Maddieson (2011). The performance of each model can be compared by Akaike Information Criteria (AIC), which measures the performance of a model while penalizing its complexity. The pR^2 (McFadden's pseudo-R squared) value indicates how much variation is explained by adding the 'neighbor' factor; large pR^2 for clicks means that whether a language has clicks or not can largely be accounted for by its neighbors. Small pR^2 for th-sounds, on the other hand, means that whether neighbors have th-sounds has little predictive power. This suggests that, although both clicks and th-sounds are uncommon features of a phoneme system, the rarity of th-sounds is more likely to reflect its linguistic (dis)preference, while the rarity of clicks may be a historical accident.

3. Implicational universals. The same approach can also be applied to implicational universals. For example, consider the hypothesis that the VO basic word order implies that

Table 2: Autologistic regression results

Clicks

	AIC	pR^2
Click $\sim$ 1	90.6	
Click $\sim$ neighbor	28.7	72.1%

Th-sounds

	AIC	pR^2
Th-sound $\sim$ 1	306.5	
Th-sound $\sim$ neighbor	303.5	1.6%

the relative clause is postposed. We can test this hypothesis with by adding the order of verb and object (henceforth V/O):

$$\text{N/Rel} = \alpha + \beta_1 \text{V/O} + \beta_2 N + \beta_3 N * \text{V/O} + \epsilon$$

Table 3 shows our results. We can see that including both the neighbor factor and V/O significantly improves the model, compared to the models only with the neighbor factor or V/O. This suggests that the VO word order does predict that the relative clause is likely to be postposed, even when the neighbor factor is taken into account.

Table 3: An analysis of an implicational universal

	AIC	pR^2
N/Rel $\sim$ 1	678.9	
N/Rel $\sim$ neighbor	322.7	52.9%
N/Rel $\sim$ V/O	396.4	42.0%
N/Rel $\sim$ neighbor + V/O	254.9	63.2%
N/Rel $\sim$ neighbor + V/O + neighbor * V/O	256.4	63.3%

4. Further issues. A number of problems remain in this approach. First, this approach does not solve the lack of random sampling. Second, this approach does not distinguish genealogical factors from language-contact factors. While it is straightforward to add the genealogical factor to the model, we must be cautious about the issue of collinearity because the genealogical and language-contact factors are expected to highly correlate with each

other. Third, the method of autologistic regression itself is not without criticisms, and more sophisticated statistical models have been proposed (Dormann, 2007).

References

Bickel, B. (2008). *A general method for the statistical evaluation of typological distributions.* ms., University of Leipzig.

Dormann, C. F. (2007, October). Assessing the validity of autologistic regression. *Ecological Modelling*, *207*(2-4), 234–242.

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

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, *68*(1), 81–138.

Dryer, M. S. (2011). Relationship between the order of object and verb and the order of relative clause and noun. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library.

Dryer, M. S., & Haspelmath, M. (2011). *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.

Hawkins, J. A. (1994). *Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Maddieson, I. (2011). Presence of uncommon consonants. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library.

Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology*, *4*(3), 307–333.