Further analysis suggesting that dry air impacts the rate at which languages rely on vowels

Caleb Everett
caleb@miami.edu

1.Introduction

This paper is an informal addendum to “Languages in Drier Climates Use Fewer Vowels”. In that paper I offer evidence for an association between vowel usage and humidity. In this addendum I offer a few additional analyses, with some different methods and an alternate database. With respect to the former, here I present results obtained with linear mixed effects methods. With respect to the latter, I include analyses based not only on the WALS genealogical classification of languages but also on the Autotyp database’s classification. (Bickel & Nichols 2017) The latter classification was employed to ensure that the findings in the main paper were not overly reliant on one particular categorization of the world’s language families. The vowel index data for the 4012 word lists used in the main paper were cross-referenced with the Autotyp data, yielding 2740 word lists with Autotyp classifications.

I’m going to report the results here informally. If you are interested in any details, all data, code, and more verbose results are available via email.

2. Mixed effects tests

The first results worth discussing are those obtained with the mixed effects model. (These tests utilized the *lme4* package in R.) With this approach, humidity is treated as a fixed effect while language family and geographic region are treated as random effects. In the case of the Autotyp database, there are 299 linguistic “stocks”, as opposed to the 229 WALS families used in the main paper. The autotyp database also groups languages according to ten geographic regions. I used the regions native to that database to classify the language varieties according to the 6 continent-level regions used in the main paper. The mixed effects method allows for the contrast of models in which humidity is treated as a fixed effect with those in which it is not included as an effect (the latter are the null models). The models including humidity as a fixed effect were contrasted with the relevant null models in a pairwise fashion. For all of the contrasts listed in a)-d), a significant disparity was obtained between the “humidity” model and the null model. This significant disparity held for both random intercept contrasts and random slope contrasts. (*p* values ranged from 0.0001 to 0.02, and were generally lower for the WALS-based tests.)

a) Tests with vowel index as the dependent variable, with and without humidity as a fixed effect. The 6 major regions and 229 WALS families were treated as random effects.
b) Tests with vowel index as the dependent variable, with and without humidity as a fixed effect. The 6 major regions and 299 Autotyp stocks were treated as random effects.

c) Tests with alternate vowel index (see methods section of main paper) as the dependent variable, with and without humidity as a fixed effect. The 6 major regions and 229 WALS families were treated as random effects.
d) Tests with alternate vowel index as the dependent variable, with and without humidity as a fixed effect. The 6 major regions and 299 Autotyp stocks were treated as random effects.

In short, the series of tests with mixed effects models support the idea that humidity is a predictor of vowel usage, even after controlling for region and phylogeny in this manner.

3. Cross-family and cross-region regressions employing Autotyp classification

As we see in Figure A, the same pattern found in Figure 3 of the main paper, for WALS families, is also observed when we rely on the Autotyp stocks.

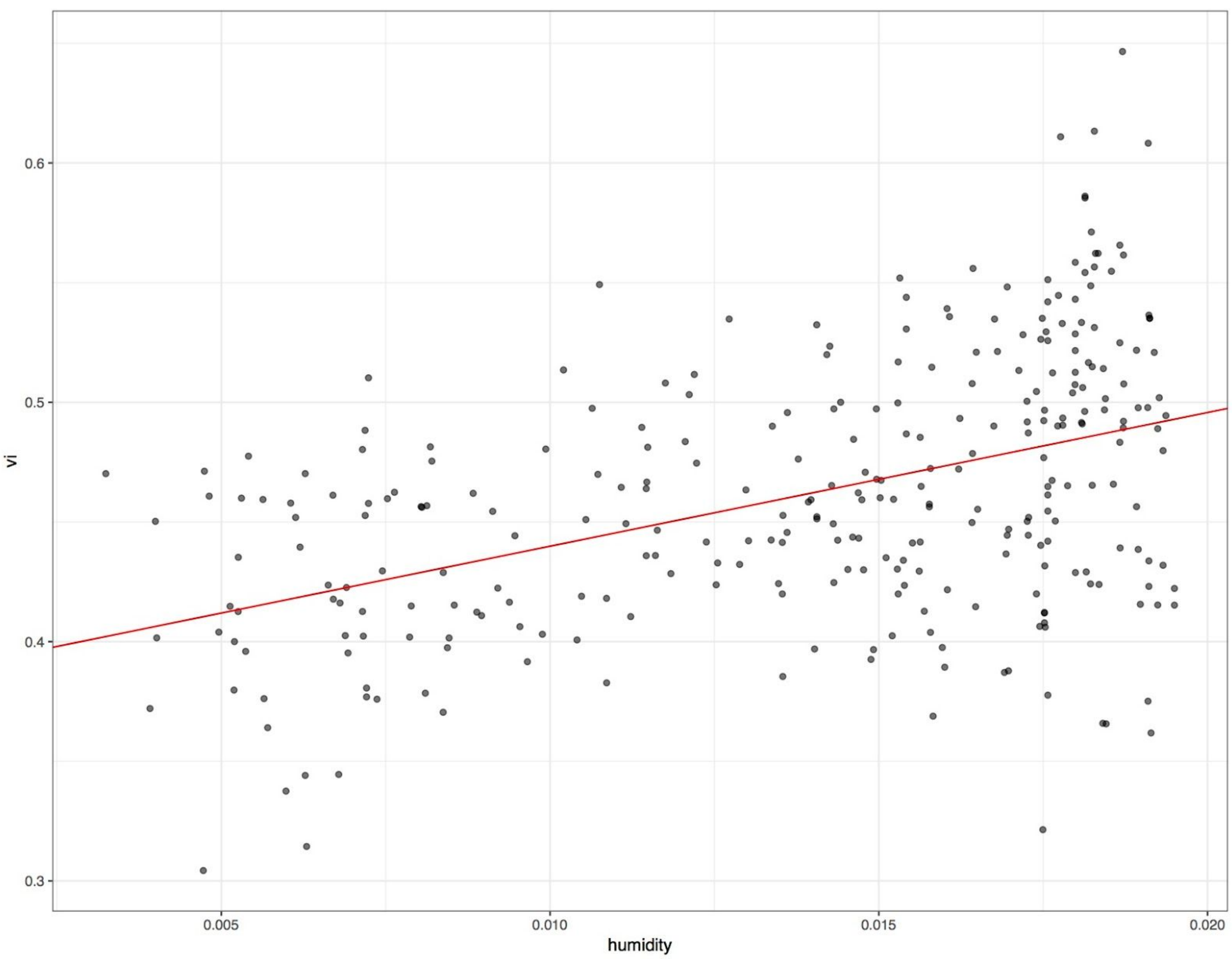

Figure A. Distribution of mean vowel indices and mean humidity values obtained for 299 Autotyp stocks.

The association in Figure A is significant according to a beta regression between mean family vowel index and mean family humidity. (*Pseudo* R^2=0.2, p=0.000) A regression between median family vowel index and median family humidity reveals similar results. (*Pseudo* R^2=0.189, p=0.000) The *Pseudo* R^2 value for the beta regression of the mean alternate vowel index with mean humidity is 0.186, while that for median alternate vowel index with median humidity is 0.173. In both cases, the association is again highly significant (p=0.000).

Turning now to median and mean values of regions, we again find a robust cross-regional pattern like that observed in Figure 5 of the main paper. For a slightly different perspective, I rely here on the 10 regions of the Autotyp database.

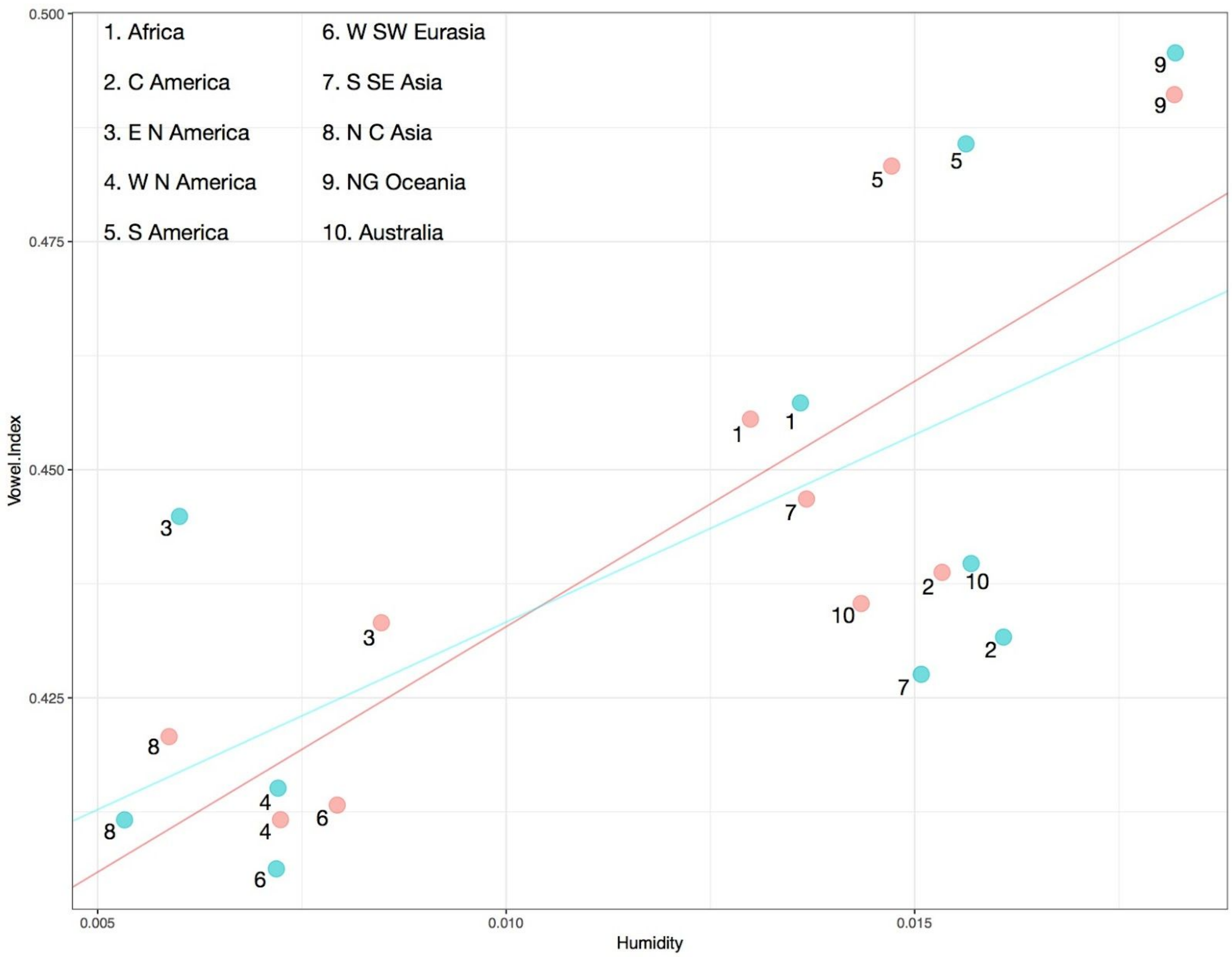

Figure B. Mean/median vowel index and humidity values for 10 autotyp regions, based on autotyp stocks. (Means are in red, medians in cyan.) For each region, the mean values of the stocks are obtained, and then the mean of those means is used to represent the given region. Additionally, the medians of the stocks are obtained, and then the median of those stock medians is used to represent the given region. The *Pseudo* R^2 value for the means in Figure B is 0.678 (p=0.004). For the medians, it is 0.452 (p=0.025).

4. Phylogenetically controlled within-region tests, using Autotyp instead of WALS classification

For each of the five landmasses, within-region beta regressions were run. These were based on the mean and median values for each stock. The association between humidity and vowel index revealed itself on each continent, with the exception (again) of Australia. Interestingly, though, the Autotyp data (which categorize Australian languages into numerous stocks) do not reveal a negative association between humidity and vowel index as in the main paper's analysis. Instead, there is no association of any sort between stock-based median/mean vowel indices and humidity values in Australia. Below are the values obtained for each of the four major continents, for the beta regressions of the mean and median values, respectively.

	Means		Medians	
	Pseudo R^2	*p*	Pseudo R^2	*p*
Africa	0.233	0.000137	0.181	0.00112
Eurasia	0.223	0.00199	0.196	0.00499
South America	0.113	0.003359	0.123	0.002116
North America	0.062	0.0592	0.048	0.0986

As with the main analysis, a positive association between vowel index and humidity was observed within each of the four principal landmasses. Once again, the association showed itself most weakly in North America. The association in North America is slightly weaker for the Autotyp classification than in the case of the WALS classification. On the other hand, the negative association observed in the main paper, vis-a-vis Australia, is not observed in the Autotyp data. (The interaction between mean vowel index and mean humidity across 26 Autotyp stocks in Australia is marginally positive, with a negligible *Pseudo* R^2 of 0.0001183.)

5. Conclusion

The findings in this brief and informal addendum suggest that the association between humidity and vowel usage is evident with mixed methods approaches, and also evident when an alternate classification of languages is utilized. As I stressed in main paper, the fact that this association exists does not demonstrate that ambient air characteristics have an impact on the development of languages. However, for the reasons adduced there and in my previous work with D. Blasi and S. Roberts, and given the robustness of this newly uncovered association, I think the possibility of such an impact merits consideration now more than ever.