Linguistic niches emerge from pressures at multiple timescales

Molly Lewis
mll@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

What accounts for the vast diversity in the world's languages? We explore one possibility: languages adapt to their linguistic environment (*Linguistic Niche Hypothesis*; Lupyan & Dale, 2010). Recent studies have found support for this hypothesis through correlations between aspects of the environment and linguistic structure. We synthesize this previous work and find that languages spoken in cold, small regions tend to be more complex across a range of linguistic features. We also test a novel prediction of the Linguistic Niche Hypothesis by examining the learnability of languages for L1 speakers.

Keywords: Linguistic Niche Hypothesis; language evolution

Introduction

What factors shape language? Psychologists have made significant progress understanding this question in the domains of communicative interaction and children's developmental trajectories. In both cases, accounts rely on positing two pressures on the cognitive system—one internal and one external. In the case of communication, theorists argue that speakers are influenced by cognitive constraints (minimize effort) and by the needs of the communicative partner (be understandable; Horn, 1984). In the case of acquisition, there are internal maturational constraints, as well as external pressures from the quality and quantity of linguistic input (Hart & Risley, 1995). In the present paper, we explore the possibility that the same two pressures—system internal and external—may also shape *language systems*.

Central to this hypothesis is the notion of a timescale: there are different units of time over which processes operate, and processes at shorter timescales influence those at longer timescales (Blythe, 2015, see also Fig. 1). At the shortest timescale are individual utterances in communicative interactions (pragmatics). At a longer timescale is language acquisition. Both experimental and modeling work suggest that communicative interactions at the pragmatic timescale influence processes like word learning at the acquisition timescale (e.g., Baldwin, 1991; McMurray, Horst, & Samuelson, 2012; Frank, Goodman, & Tenenbaum, 2009; Frank & Goodman, 2014).

A third relevant timescale is language evolution: the timescale over which entire language systems change. As for acquisition, there is evidence that language systems may be the product of processes at the pragmatic timescale. For example, languages universally structure semantic space to reflect optimal equilibria between communicative pressures (e.g., Kemp & Regier, 2012; Regier, Kay, & Khetarpal, 2007; Baddeley & Attewell, 2009).

However, the presence of communicative pressures at the pragmatic timescale is unable to explain cross-linguistic variability in linguistic structure. That is, why does Polish have rich morphology but English relatively sparse? A growing body of work argues that this variability may be due to cognitive constraints internal to the language learner (Chater & Christiansen, 2010) as well as properties of the environmental context (Nettle, 2012). This hypothesis, termed the *Linguistic Niche Hypothesis* (Lupyan & Dale, 2010; Wray & Grace, 2007), suggests that language systems adapt to the internal and external pressures of the linguistic environment.

A number of recent studies provide correlational support for this proposal. At the lowest level of the linguistic hierarchy, languages with larger populations are claimed to have larger phonemic inventories (Atkinson, 2011; Hay & Bauer, 2007), but shorter words (Wichmann, Rama, & Holman, 2011). Speakers with more second language learners have also been suggested to have fewer lexical items (Bentz, Verkerk, Kiela, Hill, & Buttery, 2015). At the level of morphology, speakers with larger populations tend to have simpler morphology (Lupyan & Dale, 2010; Bentz & Winter, 2013). Finally, there is also evidence that population size may influence the mappings between form and meaning. In particular, this work suggests that languages tend to map longer words to more complex meanings (Lewis, Sugarman, & Frank, 2014), but that this bias is smaller for languages with larger populations (Lewis & Frank, 2016).

The plausibility of the Linguistic Niche Hypothesis depends largely on the presence of a possible mechanism linking environmental features to aspects of language systems. A range of proposals have been suggested (Nettle, 2012). For example, one possibility is that children (L1) and adult (L2)

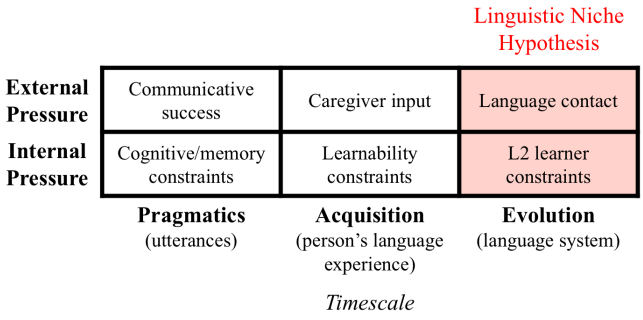

Figure 1: Pressures on language internal and external to the cognitive system, at three different timescales. The Linguistic Niche Hypothesis suggests that language evolution is influenced by the internal and external pressures in the particular environmental context in which a language is spoken.

language-learners differ in their learning constraints. In particular, children may be better at acquiring complex morphology than adults, and so languages with mostly children learners may tend to have more complex morphology. A second possibility is that speakers in less dense social networks have less variable linguistic input, and this leads the language system to have more complex morphology.

Providing evidence for these mechanisms is empirically challenging, however. Because there are many factors that shape a linguistic system, large datasets are needed to detect a correlation with environmental factors. In addition, there is non-independence across languages due to genetic relationships and language contact, and so data from a wide range of languages are needed to control for these moderators (Jaeger, Graff, Croft, & Pontillo, 2011). Third, the hypothesized mechanisms are somewhat underspecified, and the dynamics between these different factors may be complex, trading-off with each other in non-obvious ways (e.g., Wichmann et al., 2011). Finally, the large scale of this hypothesis makes it difficult to directly intervene on, and so we must rely primarily on correlational data to make inferences about mechanism.

In this work, we try to address some of these challenges by clarifying the empirical landscape. We do this by aggregating across datasets that find covariation between environmental variables and linguistic structure. This serves two purposes. First, it allows us to examine the relationship between the same set of environmental predictors across a range of linguistic features. And, second, it allows for the same analytical techniques and areal controls to be used across datasets. By addressing these inconsistencies, we are better able to directly compare relationships between environmental and linguistic features. A more coherent picture of the empirical landscape may in turn provide insight into the mechanism linking language systems to their environments.

We also directly explore the link between the acquisition and language evolution timescales by testing a novel prediction about the relationship between L1 and L2 learnability. If the correlation between environments and linguistic complexity is due to different learning constraints of L1 and L2 learners, then we should expect different kinds of languages to favor acquisition for different populations. In particular, we explore the prediction that languages that are more easily learnable by L2 populations are harder to learn for L1 learners.

In what follows, we first present a study examining the relationship between environmental and linguistic features using the same analytical techniques (Study 1). In Study 2, we examine the relationship between cross-linguistic variability in mean age of acquisition and environmental and linguistic features.

Study 1: Environmental pressures on language systems

The hypothesis of interest suggests a relationship between environmental and linguistic features, though the direction and magnitude of this relationship varies across the previous literature. To explore this variation, we combined data from five existing datasets that included environmental or linguistic data. The datasets were selected for being publicly available and containing a large sample of languages. Below we describe each of these datasets, followed by our analytical methods, and results.

Datasets

Lupyan and Dale (2010). This dataset contains grammatical information from WALS (Dryer & Haspelmath, 2013), and demographic and geographic information from Ethnologue and the Global Mapping Institute (Gordon, 2005). The demographic and geographic variables included total population of speakers, number of neighboring languages, area of region in which the language is spoken (km^2), mean and standard deviation temperature (*celsius*), and mean and standard deviation precipitation (*cm*). Using the data from WALS, we created a metric of morphosyntactic complexity calculated from 27 of the 28 morphosyntactic variables analyzed in the original paper.[1] For each variable, we coded the strategy as simple if it relied on a lexical strategy or few grammatical distinctions (e.g., 0-3 noun cases), and complex if it relied on a morphological strategy or many grammatical distinctions (e.g., more than 3 noun cases). We summed the number of complex strategies to derive a measure of morphosyntactic complexity measure for each language, including only languages with data for all 27 variables. [n = 1991 languages]

Bentz et al. (2015). Two variables were used from this dataset: ratio of L2 to L1 speakers and number of word forms. Estimates of number of word forms were taken from translations of the *Universal Declaration of Human Rights*. Number of word forms was calculated as the number of unique words divided by the number of total words (type-token ratio). Higher type-token ratio indicates more word types in that language. Speaker population data were taken from a variety of sources, where L2 speakers were restricted to adult non-native speakers only. [n = 81]

Moran, McCloy and Wright (2012). Estimates of number of consonants and vowels in each language were used from this dataset. These were originally taken from the Phoible database (Moran & Wright, n.d.). [n = 969]

Lewis and Frank (2014). This work finds that languages tend to map more complex meanings (measured via semantic norms) to longer words. The bias is estimated as the correlation (Pearson's r) between word length (in terms of number of characters) and complexity ratings for a set of 499 words translated via Google Translate. We used estimates of the correlation that partialed out the effect of spoken frequency [n = 79]

[1] WALS variable 59 was missing from the dataset.

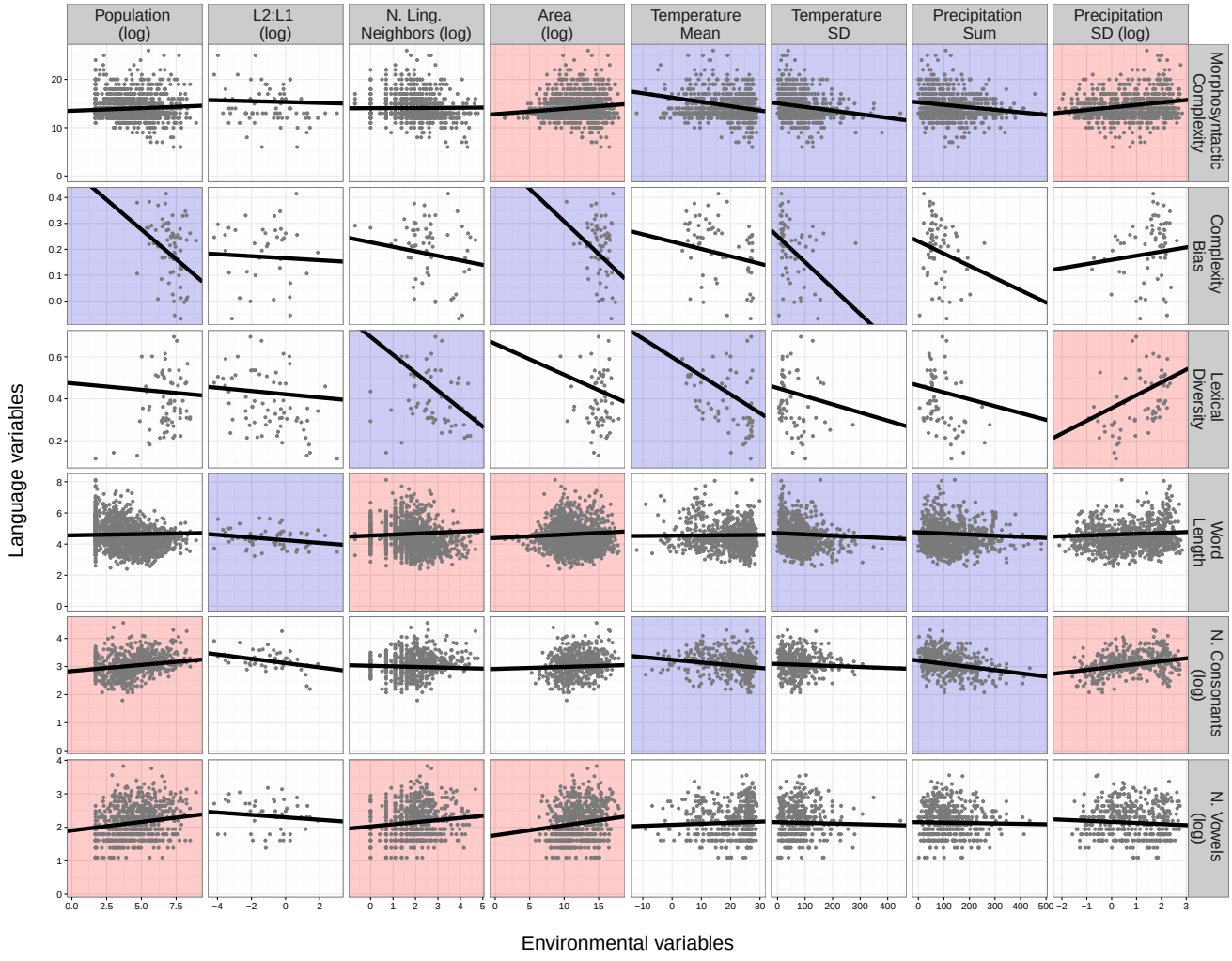

Figure 2: Relationship between environmental and linguistic variables, which each point represents a language. Red (positive) and blue (negative) indicate models where the environmental variable is a significant predictor of the linguistic variable. Lines show the fixed effect estimate (slope) and intercept of the mixed effect model. Number of languages varies across plots due to variation in the number of overlapping languages across datasets.

Wichmann, Rama, and Holman (2014). This database contains translations for 40-lexical items across many languages. Word length was calculated as the mean number of characters in the ASJPcode transcription system across words in each language. [$n = 4421$]

Aggregating across datasets, we analyzed 8 environmental variables in total: L2-L1 population ratio, total population size, number of neighbors, area of spoken region, mean and standard deviation temperature, and mean and standard deviation precipitation. These variables were selected from a larger set because they were not highly correlated with each other ($r < .8$). We analyzed 6 total linguistic variables: number of vowels, number of consonants, word length, type-token ratio, complexity bias, and morphosyntactic complexity.

Method

Datasets were merged using common ISO-639 codes. Five variables were log-transformed to better approximate a normal distribution (population, L2 to L1 ratio, number of neighbors, area, number of consonants, number of vowels).[2]

Main analysis We tested for a linear relationship between each environmental and language variable. A significant challenge in making inferences about language data is non-independence. This non-independence can come from at least two sources: genetic relatedness and language contact. Following Jaeger et al. (2011), we control for these factors statistically by using linear mixed-effects regression. We control for genetic non-independence by including a random intercept and slope by language family. We control for language contact by including country of origin as a random intercept (models with random slopes failed to converge).[3] We selected

[2]All code and data for the paper are available at `http://github.com/mllewis/langLearnVar`

[3]The model specification was as follows: `language.variable ~ environmental.variable +`

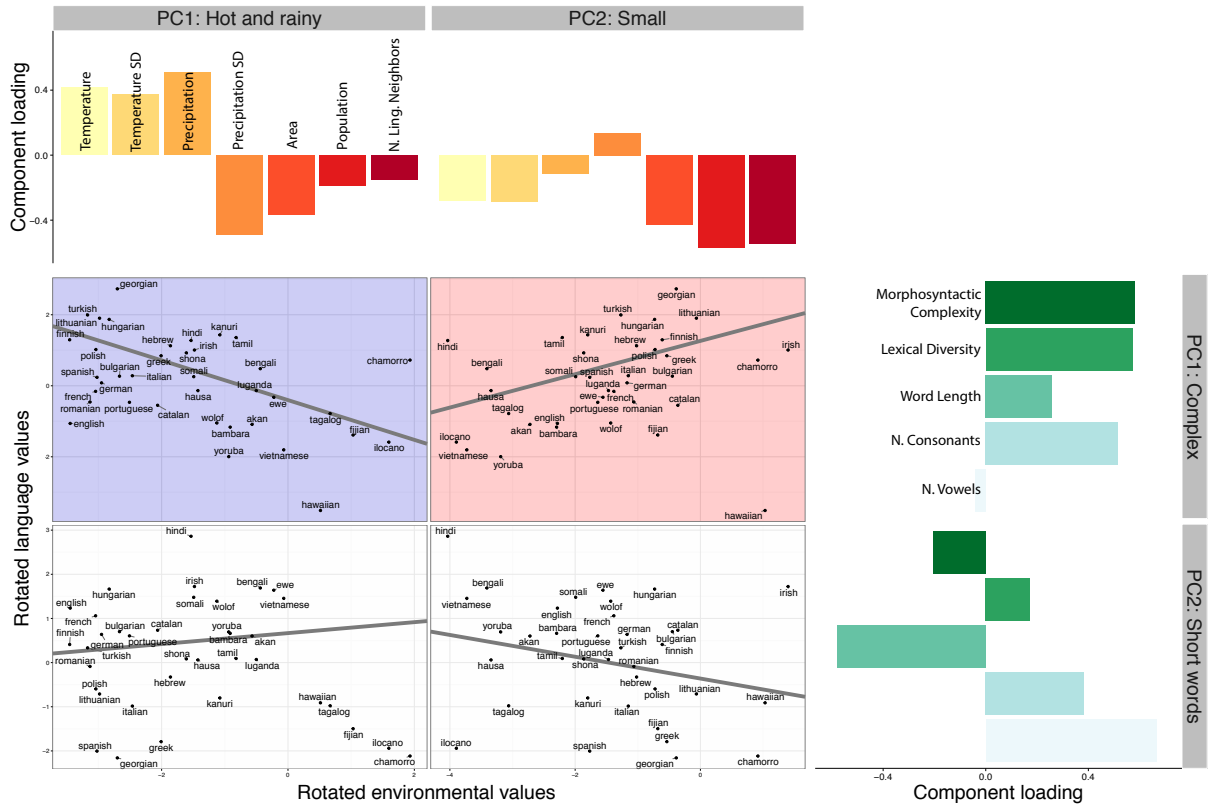

Figure 3: Languages spoken in cold, small regions tend to be more complex. The bar plots show the loadings on the first two principal components for the environmental variables (n = 7; orange) and language variables (n = 5; green). The scatter plots show the relationship between the first two principal components for both sets of variables. Each point corresponds to a language, and lines show the linear fit from the mixed effect model. Significance and direction of a linear relationship are indicated by the coloring of the scatterplot (blue: significant and negative; red: significant and positive).

country of origin as a proxy for linguistic community because it was available for all languages in our dataset. Both control variables were taken from the WALS dataset. We considered a predictor significant if the test statistic on the fixed effect coefficient exceeded 1.96.

Principal component analysis This first analysis provides a uniform analysis of the many environmental and linguistic variables that have been used to test the Linguistic Niche Hypothesis. However, the number of variables makes it difficult to distill a coherent picture from these data. Given that many of these variables are partially correlated with each other, we used a technique for reducing the dimensionality of the dataset—principal component analysis. We found the principal components associated with the variance for the environmental variables and the linguistic variables, and then fit the same model as in the primary analysis using the rotated values. Complexity bias was excluded because it was only available for a small subset of languages. All variables were scaled.

Results

In the main analysis, we fit mixed effect models predicting each language variable with each environmental variable using areal controls. The results are presented in Fig. 2. For each language variable, there was at least one environmental variable that reliably covaried, though some previously-reported effects were not significant in this analysis. We return to this in the discussion. Data can be explored interactively through an online app: `https://mlewis.shinyapps.io/lhnn/`.

The principal component analysis revealed two primary components of variance for both the environmental and linguistic variables. For the environmental variables, the first two principal components accounted for .69 of the total variance (PC1: .39; PC2: .30). The weights on these variables across the two components can be seen in the upper panel of Fig. 3. The first component loads most heavily on variables related to the climate. It can be thought of as corresponding to hot and rainy regions. The second component loads most heavily on variables related to the size of the region a language is spoken in, both in terms of number of speakers and physical size. This principal component can be roughly interpreted as the 'smallness' of a linguistic community.

`(environmental.variable | language.family) + (1 | origin.country)`.

For the linguistic variables, the first two components also accounted for most of the variance, .70 (PC1: .39; PC2: .31; right panel of Fig. 3). The first component loads positively on all variables, except number of vowels. In particular, this component is associated with more consonants, longer words, more word types, and greater morphosyntactic complexity. Broadly, this component is related to the amount of cognitive difficulty associated with learning a language. The second component is associated with having short words, but large phonemic inventories.

Figure 3 shows the relationship between the principal components. Both environmental principal components were reliable predictors of the first linguistic principal component (PC1: $\beta = -0.56$, $t = -3.52$; PC2: $\beta = 0.47$, $t = 2.08$). This suggests that languages that tend to be spoken in cold and small regions are more likely to be more complex. Neither of the environmental principal components were reliable predictors of the second linguistic principal component.

Discussion

These two analyses suggest that more complex languages are spoken in cold, small regions. Importantly, we find this relationship across a range of linguistic features—morphosyntactic complexity, linguistic diversity, word length, and consonant inventory—using the same analytic technique across all measures.

This finding is broadly consistent with previous work that finds relationships between individual metrics of complexity and various demographic variables. Nevertheless, we find null effects for several reported relationships in the literature. For example, the relationship between population size and morphosyntactic complexity (Lupyan & Dale, 2010) is not reliable in our model, though the correlation is significant (r =.08; p <.001) and we replicate their finding in a binned analysis (Fig. 3 of Lupyan & Dale, 2010). There are many possible reasons for these differences (e.g., different measure of complexity, different areal controls), highlighting the need for a common analytical approach across datasets.

Why might languages in small, cold regions have more complex languages? One possible mechanism is that languages spoken in larger places have more L2 learners, and that L2 learners are less skilled than L1 learners at acquiring complex language. As a result, these languages adapt by simplifying. The relationship between climate and linguistic complexity is less clear, but a possibility is that speakers in colder regions are less itinerant, and therefore have less contact with adult speakers of other languages.

Study 2: Variability in L1 learning

The proposed mechanism in Study 1 makes an important assumption: L2 learners, but not L1 learners, are poor learners of linguistic complexity. Lupyan and Dale (2015) have argued that morphological complexity in fact *facilitates* learning for L1 learners by providing redundancy in the linguistic signal. A straightforward prediction of this hypothesis is that languages that are more easily learnable by L2 learners will be less learnable by L1 learners.

In Study 2, we explore this prediction. As a proxy for language learnability for L1 learners, we use the mean age of acquisition (AoA) of words in a language by L1 learners (children). If there is the predicted tradeoff between learnability for L1 and L2 learners, languages that are more complex should have earlier AoAs.

Method

We use subjective measures of AoA from the the Łuniewska et al. (2015) norms. These AoAs were collected from adult participants for the translation equivalents of 299 words in 25 languages. To evaluate the validity of this measure, we compared these ratings to more objective measures of AoA collected from parent-report using the CDI (Wordbank; Frank, Braginsky, Yurovsky, & Marchman, in press). We fit a model predicting the objective ratings with the subjective ratings for the small sample of common languages ($n = 7$). We included language as a random by-intercept and by-slope effect. Subjective ratings were a strong predictor of objective ratings ($\beta = 1.00$, $t = 5.45$), suggesting that the Łuniewska et al. (2015) norms were a reasonable proxy for cross-linguistic AoA.

We averaged across words in the Łuniewska et al. (2015) database to get a mean AoA for each language. We then used the same mixed-effect model as in Study 1 to predict AoAs with each of the linguistic and environmental variables analyzed in Study 1.

Results

Number of consonants was a reliable predictor AoA ($\beta = 1.04$, $t = 1.97$). In addition, temperature positively predicted AoA ($\beta = .13$, $t = 2.57$; $\beta = 1.04$, $t = 1.97$) and variability in precipitation negatively predicted AoA ($\beta = -1.6$, $t = -2.83$). No other linguistic or environmental variables were significant predictors of AoA.

Discussion

Study 2 explores a prediction about a mechanism for the relationship between population size and linguistic complexity: L1 learning is facilitated by complexity in the linguistic signal (via redundancy), but L2 learners are hindered. We find only limited support for this proposal in the current analysis. Of the factors that loaded on "complexity" in principal component analysis, number of consonants was the only reliable predictor of AoA.

Nevertheless, we do find several surprising correlates of AoA—number of consonants, temperatures, and precipitation variability—even in this very small sample of languages. The mechanism underlying these relationship is not clear. It could be for example that L1 learners in colder populations have more language input, and therefore earlier AoAs. Or, if we assume that temperature and variability in precipitation are proxies from L2 pressure (based on Study 1), it could be that languages with more L2 pressures have later AoAs, and

therefore are harder for L1 learners to acquire. A larger sample of languages will be needed to address these questions.

Conclusion

Languages vary in many ways across multiple timescales of analysis. Here we suggest that this variability can be accounted for by considering the relationship between these timescales and two types of pressures, those internal and external to the cognitive system. In the present work, we have explored a hypothesis at the language evolution timescale—the Linguistic Niche Hypothesis—which suggests that cross-linguistic variability is the result of different cognitive constraints of learners and environmental pressures.

We contribute to the empirical findings related to this hypothesis by synthesizing previous correlational evidence using common analytical techniques across datasets, and by testing a critical novel prediction of this hypothesis at the language acquisition timescale. Across a range of linguistic and environmental metrics, we find that more complex languages tend to be spoken in smaller, colder regions. We also find evidence evidence that L1 learning variability may be related to aspects of the language and the environment.

Accounting for variability at the timescale of language evolution is an empirically challenging enterprise. Moving forward, we suggest that a fruitful avenue for progress is holistic descriptions of the empirical landscape, and appeals to processes at multiple timescales of analysis.

Acknowledgments

We would like to thank Gary Lupyan and Rick Dale for sharing their data with us.

References

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*, 346–349.

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment information theory shows why we have only three lightness terms. *Psychological Science*, *20*, 1100–1107.

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*, 874–890.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers have fewer word forms.

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, *3*, 1–27.

Blythe, R. A. (2015). Hierarchy of scales in language dynamics.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*, 1131–1157.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://wals.info/`

Frank, M., Braginsky, M., Yurovsky, D., & Marchman, V. A. (in press). Wordbank: An open repository for developmental vocabulary data.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578.

Frank, M., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Gordon, R. (2005). Ethnologue: Languages of the world. SIL International.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *83*, 388–400.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context*, *42*.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*, 281–320.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Lewis, M., & Frank, M. C. (2016). Learnability pressures influence the encoding of information density in the lexicon learm. In *The evolution of language: Proceedings of the 11th international conference.*

Lewis, M., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowsk, B., Southwood, F., Pomiechowska, A., ... others (2015). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*.

Lupyan, G., & Dale, R. (2015). The role of adaptation in understanding linguistic diversity. *The Shaping of Language: The Relationship between the Structures of Languages and their Social, Cultural, Historical, and Natural Environments*.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831.

Moran, S., & Wright, R. (n.d.). *Phonetics information base and lexicon (phoible).* Retrieved 2009, from `http://phoible.org`

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1829–1836.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436–1441.

Wichmann, S., Rama, T., & Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, *15*, 177–197.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*, 543–578.