Language pressures across multiple timescales

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

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

Abstract

Keywords: Linguistic Niche Hypothesis; language evolution

Introduction

Why does language vary? Psychologists have made significant progress understanding linguistic variability 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 a speaker is influenced by cognitive constraints (minimize effort) and by the needs of the communicative partner (get your message across; Horn, 1984). In the case of acquisition, there are internal maturational constraints [cite], 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 account for variability in *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, 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; M. Frank, Goodman, & Tenenbaum, 2009; M. C. 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, which has been 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.

There are a number of pieces of evidence that environmental factors may indeed shape language systems. 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, evidence suggests that 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)

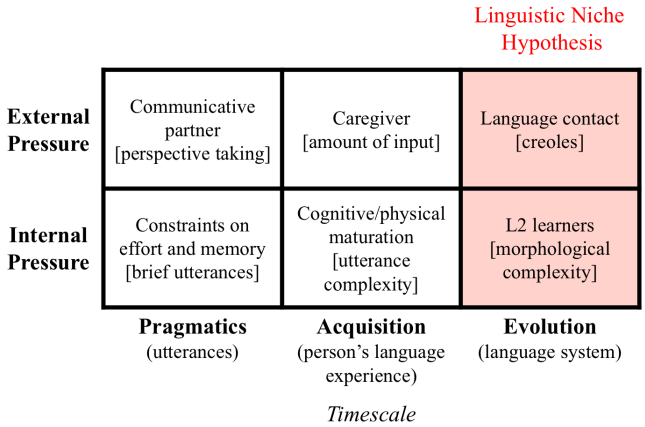

Figure 1: Pressures on language internal and external to the cognitive system, at three different timescales. Brackets show examples of phenomena that result from a particular pressure at a particular timescale. The Linguistic Niche Hypothesis suggests that language evolution is influenced by the internal 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.

Testing these mechanisms is empirically challenging, however. Because there are many factors that shape a linguistic system, large datasets are needed to detect an correlation with environmental factors. In addition, many languages are non-independent because of genetic relationships and language contact, and so data from a wide range of languages are needed to control for these moderators. Third, the scale of this hypothesis makes it difficult to directly intervene on the mechanism. Finally, the hypothesized mechanisms are somewhat underspecified, and the dynamics of these different factors may be complex, trading-off with each other in non-obvious ways (e.g. Wichmann et al., 2011).

In this work, we try to address 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. Importantly, a more coherent picture of the empirical landscape may provide insight into the mechanism linking language systems to their environments.

We also more directly address the question of mechanism by examining variability in the mean age of acquisition of words for L1 learners across languages. Evidence that this variability is related to an aspect of the linguistic system (such as number of phonemes) would suggest that L1 learners, and not L2 learners, are the relevant environmental factor shaping that aspect of the linguistic system.

In what follows, we first present a set of analyses examining the relationship between environmental and linguistic features using the same analytical techniques. We then examine the relationship between mean age of acquisition in a language and aspects of the linguistic system.

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 method, 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; *Seamless Digital Chart of the World*, n.d.). 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*). The metric of morphological complexity was calculated from 27 of the 28 morphosyntactic variables[1] analyzed in the original paper. 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]

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 *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]

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 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. We analyzed 6 total linguistic variables: number of vowels, number of consonants, word length, type-token ratio, complexity bias, and morphological complexity.

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

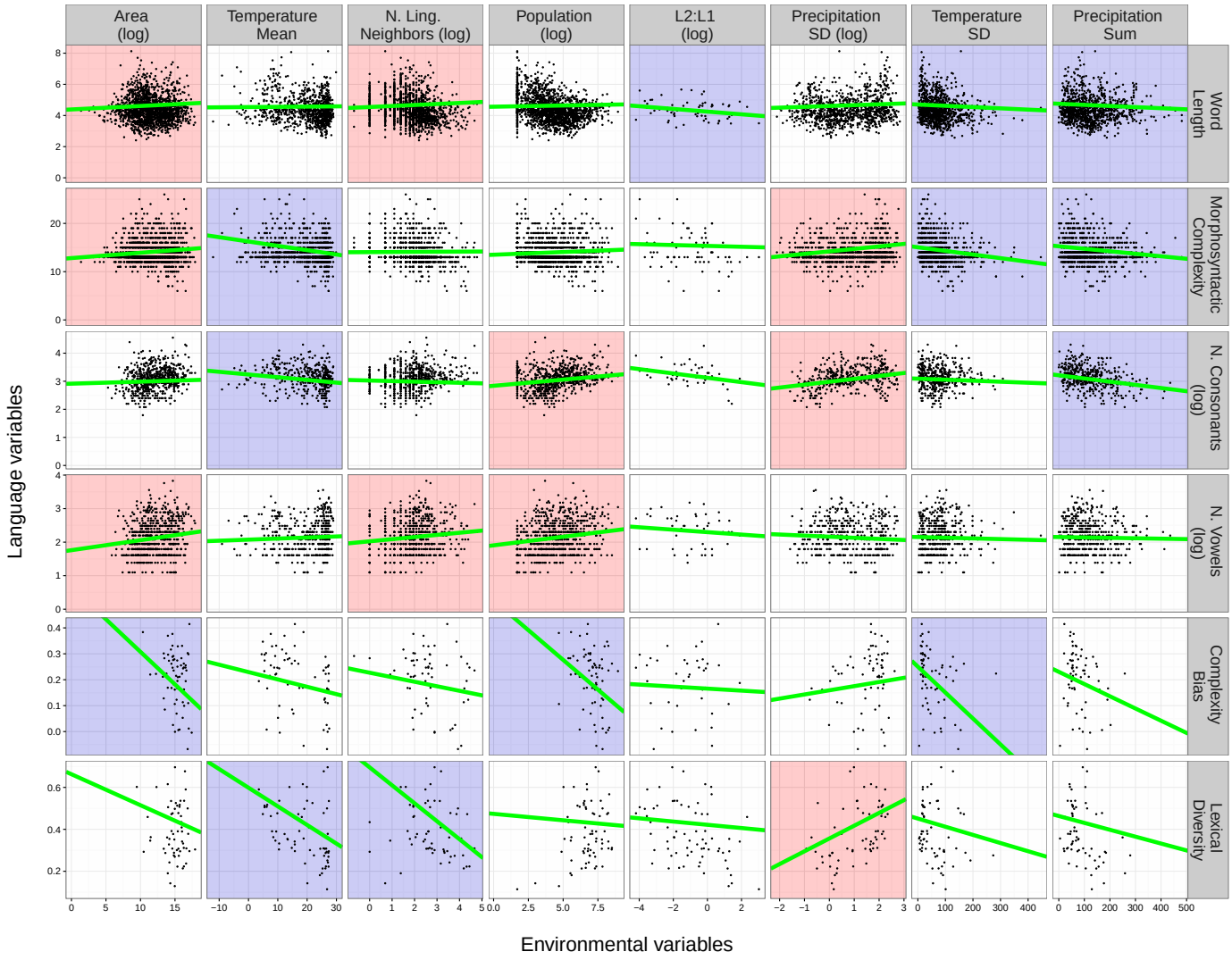

Figure 2: Relationship between environmental and linguistic variables, which each point represents a language. Red (positive) and blue (negative) indicate models where the environmental variable is a significant predictor of the linguistic variable. Linear fits are from the fixed effect estimate of the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

Method

We test 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, Graff, Croft, and Pontillo (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). While not ideal, we selected country of origin as a proxy for linguistic community because it was available for all languages in our dataset. Both language family and country of origin were taken from the WALS dataset.[2]

Results

In our first analysis, we fit mixed effects models predicting language variables with environmental factors using areal controls. In Analysis 2, we reduce the dimensionality of our data using principle component analysis, and then fit the same models as in Analysis 1.

Analysis 1: Relationship between environmental pressures and language systems with areal controls. We log-transformed five of our variables to better approximate a normal distribution (population, L2 to L1 ratio, number of neigh-

[2]The model specification was as follows: `language.variable ~ environmental.variable + (environmental.variable | language.family) + (1 | origin.country)`.

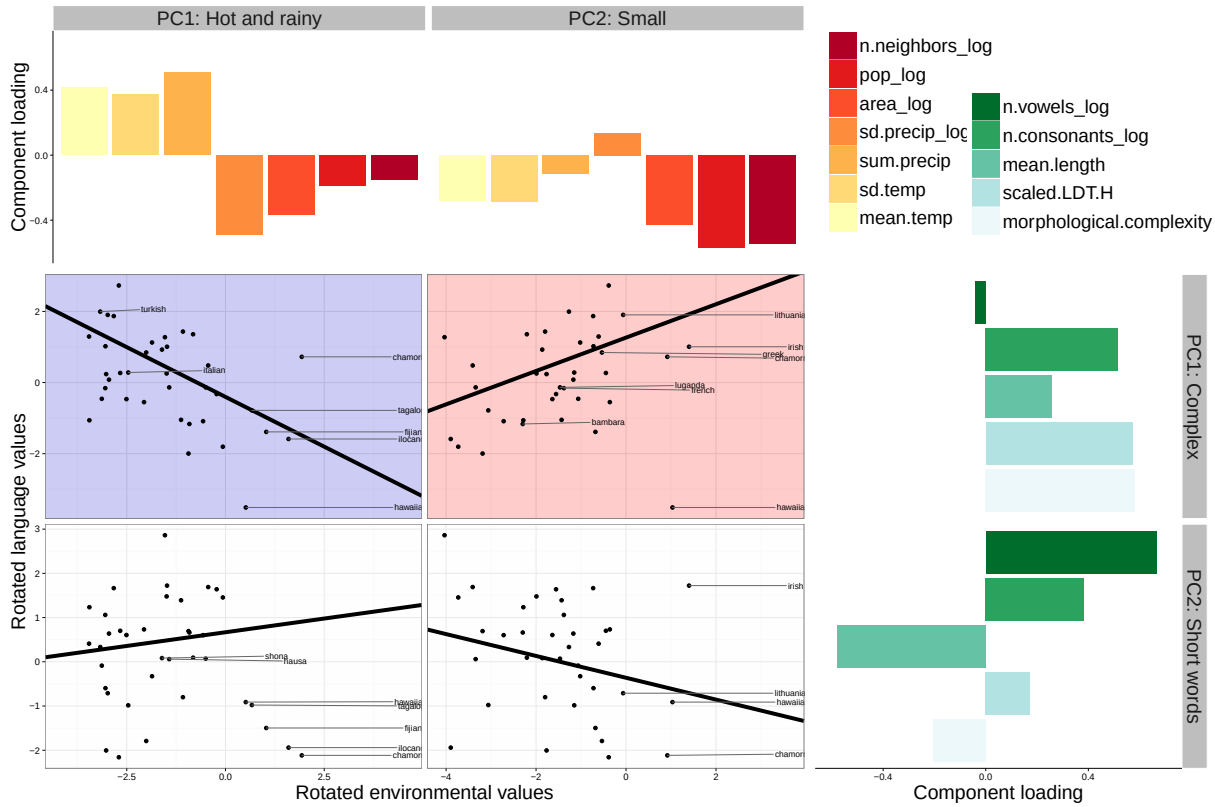

Figure 3: Languages spoken in small, cold regions tend to be more complex. The barplots show the loadings on the first two principle components for the environmental variables (n = 7; orange) and language variables (n = 5; green). The scatter plots show the relationship between the first two principle 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 on the scatterplot (red: significant and positive; blue: significant and negative).

bors, area, number of consonants, number of vowels) [3]. We then fit mixed effect models predicting each language variable with each environmental variable. We considered a relationship significant if the test statistic on the fixed effect coefficient exceeded 1.96. Fig. 1 summarizes the results. [Discussion of findings]. Data can be explored interactively through an online app: [here].

Analysis 2: Principle component analysis. Analysis 1 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 considered 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—principle component analysis. We found the principle components associated with the variance for the environmental variables and the linguistic variables (actually a subset), and then examined the relationship between the primary principle components.

All variables were first scaled. For the environmental variables, the first two principle 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. 2. The first component weights most heavily on variables related to the climate. It can be thought of as corresponding to hot and wet regions. The second component weights 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 principle component can be roughly interpreted as the 'smallness' of a linguistic community.

For the linguistic variables, the first two components also accounted for most of the variance, .70 (PC1: .39; PC2: .31; right panel of Fig. 2). The first component weights 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.

We then fit the same model as in Analysis 1, using the rotated values for the first two principle components for the environmental and linguistic variables. The plots in Fig. 2, show

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

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

Discussion

L1 learning and language systems

(Łuniewska et al., 2015)

Discussion and Conclusion

* stuff about informative about the origins of language

Other work has empirical + modeling Some attempts (Silvey, Kirby, & Smith, 2015) (Perfors & Navarro, 2011)

(Wichmann et al., 2011) (?, ?) (Smith & Wonnacott, 2010) (Slobin & Bever, 1982)

(Sapir, 1912) (REALI, CHATER, & CHRISTIANSEN, 2014)

(Lupyan & Dale, n.d.)(Lupyan & Dale, 2010) (Kirby, Cornish, & Smith, 2008) Meaning. (Silvey et al., 2015) (Perfors & Navarro, 2011) Critically,

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. *arXiv preprint arXiv:1505.00122*.

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

Frank, M. C., & 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.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*, 10681–10686.

Lewis, M., & Frank, M. C. (2016). Learnability pressures influence the encoding of information density in the lexicon learm. In *The evolution of language 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. (n.d.). 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*.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*, e8559.

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.

Perfors, A., & Navarro, D. (2011). Language evolution is shaped by the structure of the world: An iterated learning analysis. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 477–482).

REALI, F., CHATER, N., & CHRISTIANSEN, M. H. (2014). The paradox of linguistic complexity and community size. In *The evolution of language: Proceedings of the 10th international conference* (pp. 270–277).

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.

Sapir, E. (1912). Language and environment1. *American Anthropologist*, *14*, 226–242.

Seamless digital chart of the world. (n.d.). Retrieved from `http://www.gmi.org/`

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive science*, *39*, 212–226.

Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, *12*, 229–265.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*, 444–449.

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.