

Ease of Learning Explains Semantic Universals

Shane Steinert-Threlkeld and Jakub Szymanik

S.N.M.Steinert-Threlkeld@uva.nl; J.K.Szymanik@uva.nl

Institute for Logic, Language and Computation; Universiteit van Amsterdam

Abstract: Despite extraordinary differences between natural languages, linguists have identified many semantic universals – shared properties of meaning – that are yet to receive a unified explanation. Such universals are attested in various linguistic domains, including semantics of content words and function words. Recently, it has been proposed, using tools from machine learning, that universal properties of quantifiers can be explained by evoking the ease of learnability. In this paper, we show that this type of explanation can be extended to the domain of color terms. In all languages, color names denote very geometrically well-behaved (convex) regions of color space. We pursue a cognitive explanation of that fact: convexity can be explained by accounting for its role in the process of learning color terms. Therefore, we argue for a unified explanation of semantic universals, across content (color terms) and function words (quantifiers), in terms of learnability: meanings satisfying attested universals are easier to learn than those that are not. Thus, ease of learning can explain the presence of universals in many different linguistic domains.

1. Introduction

In spite of extraordinary differences between languages, linguists have identified shared properties of all languages at many levels of linguistic analysis, e.g., phonology (1), syntax (2–4), and semantics (5, 6). Linguistic universals are crucial for understanding human cognition. Because they are attested across different languages and communicative niches, they likely reflect general features of human cognitive makeup. Explaining why a universal holds requires establishing a connection between language and a feature of the mind.

So far, the literature offers very few explanatory theories of linguistic universals. Even fewer existing explanations are expressed in a precise mathematical modeling framework. A notable exception is a recent explanation in terms of optimal communication: languages vary but the scope

of possible variation is constrained by two complementary forces of simplicity and informativeness (7, 8). However, this type of explanation suffers from a serious problem: within this framework, in order to explain a given universal, one needs to come up with a complexity measure for the conceptual domain in question. Such measures can seem arbitrary, since the resulting complexity hierarchies depend on choices (e.g. of primitives) made by the theorist (9, 10).

We propose a unified cognitive explanation of linguistic universals without postulating an ad hoc complexity measure. We suggest that linguistic universals arise because expressions satisfying them are easier to learn than those that do not. To operationalize the difficulty of learning, we have not tailor-made a computational model for our specific learning tasks. Instead, we use off-the-shelf computational tools: neural network architectures currently in widespread use in deep learning (11, 12).

We focus on semantic universals in two domains: quantifiers and color terms. After presenting existing results on the former, we present new results on the latter. Our results show that universal semantic properties in both domains make meanings easier to acquire for neural networks. We have chosen these two extremely rich linguistic domains for a number of reasons. Historically, they have become a testing ground for research into the relationship between language and thought (13–15). Both domains involve the interaction of various aspects of language, cognition, and perception (16, 17). Furthermore, color terms are *content* words while quantifiers are *function* words. Our explanation of semantic universals is the first to be applied to expressions from both of these fundamental categories. A single computational explanation of the source of the semantic universals in these two domains constitutes a strong argument for the main claim of the paper that linguistic universals are the result of learnability pressure.

2. Function words: Quantifiers

The study of number terms (called here and in the linguistics literature quantifiers) has been one of the cornerstones of natural language semantics since its inception (5, 18). Quantifiers are expressed by determiners -- *every, some, most, the, a, three*, et cetera -- as they occur in syntactic

configurations like [S [NP [Det D] [N N]] [VP ...]]. Determiners express binary relations between sets of objects. For instance, the sentence *Every student is happy* is true if and only if each element in the set of students is in the set of happy things, i.e. if the former is a subset of the latter. Thus, *Every* can be taken to express the subset relation.

Only a very limited subset of all logically possible quantifiers is expressed by *simple* determiners in any natural language. Simple combinatorics tells us that there are 65,536 possible quantifiers to describe any situation consisting of two objects and two properties (19). Very few of those logically possible quantifiers are expressed in natural language. It turns out that there are strong regularities across languages in terms of which quantifiers are lexicalized. Thus, the domain of quantification provides another area of semantic universals. We now highlight two of them, showing how they are exhibited by valid inference patterns.

2.1 Quantity

The first universal captures the fact that quantifiers genuinely talk about quantities. In particular, whether or not a sentence of the form *Det CN VP* is true or not should only depend on the sizes of the sets denoted by the common noun and verb phrase (as well as their intersection and differences). To exhibit this fact, we observe the validity of the following inference pattern.

Many houses on Cambridge Ave are blue.

There are exactly as many blue and non-blue houses on El Camino Real as on Cambridge Ave.

Therefore, many houses on El Camino Real are blue.

The second premise of this inference says that the number of blue houses and of non-blue houses is equal in the two locations; the reader can verify that this pattern is valid for all choices of predicates instead of *blue* and *house on X*. This validity shows us that *many* is what we will call quantitative.

QUANTITY UNIVERSAL: All simple determiners are quantitative (18–20).

2.2 Monotonicity

For the second universal, observe that the following inference pattern is valid.

Many scientists know the R programming language.

Many scientists know a programming language.

All that we have done is replaced a more specific term -- *know the R programming language* -- with a strictly more general term -- *know a programming language*. The reader can verify that the inference pattern would be valid for any choice of expression instead of *scientist* and pair of expressions that stand in the same specific/general relation. This validity shows that *many* is upward monotone. The pattern reverses for *few*.

Few scientists know a programming language.

Few scientists know the R programming language.

This shows that *few* is downward monotone. We say that a determiner is monotone if it is either upward or downward monotone. Of course, not every quantifier is monotone: those expressed by the complex determiners *an even number of* and *at least 6 or at most 2* are not monotone.

MONOTONICITY UNIVERSAL: All simple determiners are monotone (5, 18).

3. Content words: Color Terms

The study of color naming has been on the forefront of the relativism vs. nativism debate that has driven much of linguistics, philosophy, and cognitive science in the twentieth century. This is only natural: although colors have relatively simple and well-understood structure, there is at the same time fairly complex cross-linguistic variation when it comes to color terms. At first, the debate was dominated by the simple question: which colors get names (are lexicalized) across different languages? Using their seminal World Color Survey, Kay and Berlin have proposed the existence of universal constraints on cross-language color naming in the form of a partially fixed evolutionary progression according to which languages gain color terms over time (17, 21). These are strong nonlinguistic/non-cognitive constraints that can be explained by postulating that color categories are organized around universal focal colors (22). The relativistic approach, however, seems less concerned with explaining some of the most striking universal properties of color categorization across languages (23). Apart from which colors get lexicalized, all color terms denote very geometrically well-behaved regions of color space.

CONVEXITY UNIVERSAL: All color terms denote convex regions of color space (24–27).

4. Computational Experiments and Results

4.1 Quantifier Universals

In (29) we trained a long short-term memory network to learn the meanings of quantifiers. The input was a sequence of objects, coded to which of four regions of a Venn diagram they belong to, with the two sets corresponding to the common noun and verb phrase of a *Det CN VP* sentence. For each universal, we compared one quantifier satisfying the universal with one quantifier not satisfying the universal. Results can be seen in Fig. 1.

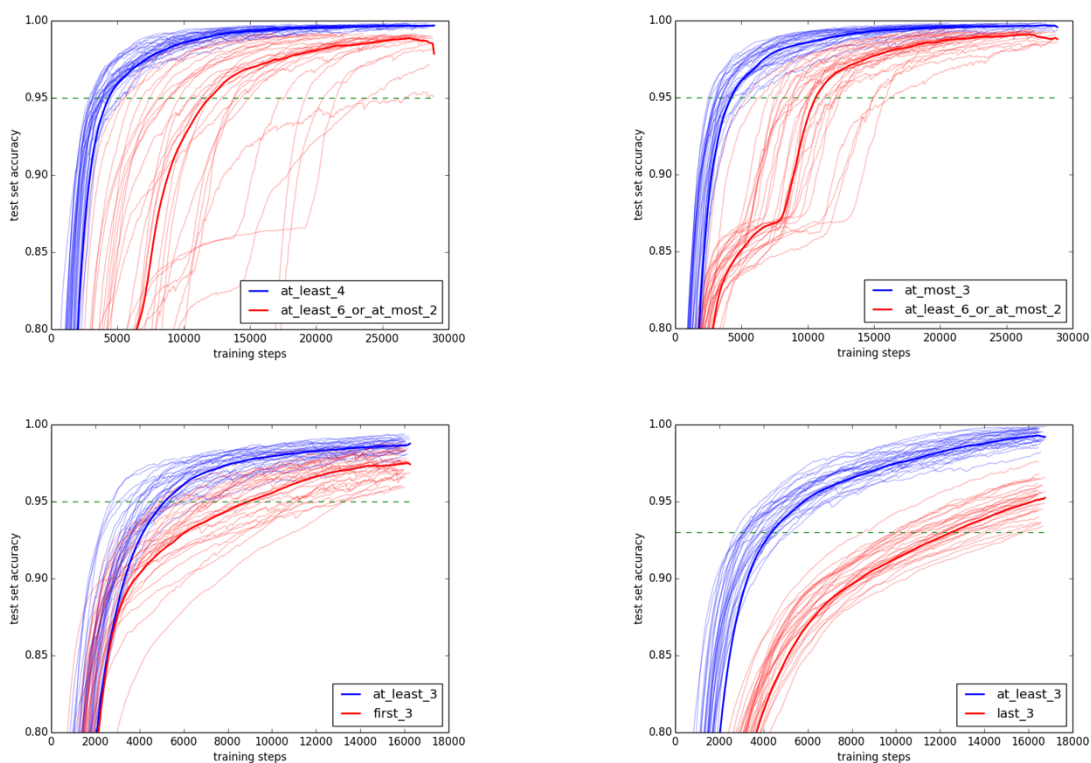


Fig. 1. Learning curves for an LSTM trained to learn quantifiers. X-axis: number of training mini-batches. Y-axis: accuracy on a held-out test set. Top row: monotonicity. Left: upward monotonicity, *at least 4* compared to *at least 6 or at most 2*. Right: downward monotonicity, *at most 4* compared to *at least 6 or at most 2*. Bottom row: Quantity. Left: *at least 3* compared to *the first 3*. Right: *at least three* compared to *the last 3*.

We ran $n=30$ trials for each universal and measured for each quantifier and each trial the *convergence point*: the first mini-batch which is above 95% test-set accuracy and for which the

future mean is above that same threshold. For each universal, we found a statistically significant difference, allowing us to conclude that the quantifier satisfying the universal converged earlier than the one that did not.

In the next section, we present new results, extending our approach to color terms.

4.2 Color Universal

We argue that the cognitive explanation in terms of learnability that we offered for function words may be the right explanation for (possibly) all semantic universals by showing that it extends naturally to content words. Specifically, we chose color terms as our case study. It has been argued that all natural language color terms denote geometrically well-behaved – specifically, convex – regions of color space (24–26). We show that convexity can be indeed explained by accounting for its role in the process of learning color terms.

4.2.1 Artificial Color Naming Systems with Varying Degrees of Convexity

Let us start by defining convexity in precise terms: Given a vector space X , a subset Y of X is convex if and only if for every two points x and y in Y , every point on the line between x and y is also in Y (i.e. the point $t*x + (1-t)*y$, for every t between 0 and 1, is in Y). Then the CONVEXITY UNIVERSAL says that the denotation of every color term is a convex region in the CIELab color space that approximates human color vision. Most importantly, the space is perceptually uniform with respect to human color vision, meaning that the distance in the space corresponds well with the visually perceived color change (25).

We generated 300 artificial color-naming systems, by partitioning the CIELab color space into distinct categories, as in Fig 2.

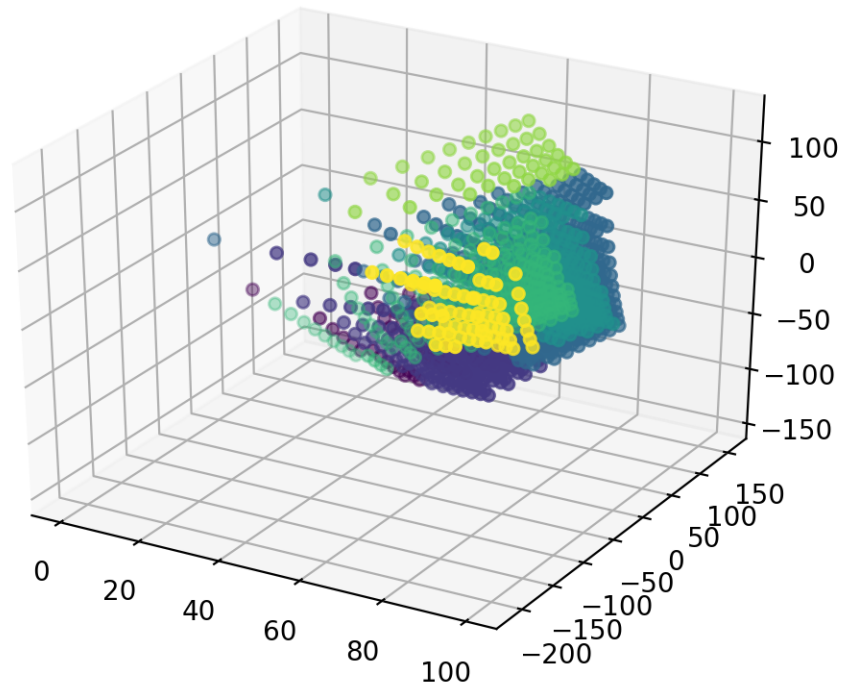


Fig. 2 Partitioning CIE-L*a*b* Space

The obtained color systems varied in the *degree of convexity*, measured as the average area of the convex hull of each color that is covered by that color.

We used the following algorithm for generating artificial color naming systems:

Parameters: temperature (t), connectedness (c), initial ball size (b)

Inputs: a set of points X , distance measure D , number of categories N

Step 1: generate initial labeling

1. Unlabeled := X ; Labeled _{i} = emptyset for $i=1, \dots, N$
2. Choose x_1, \dots, x_N randomly

1. Assign x_i to category i (i.e. add x_i to Labeled _{i} and remove it from Unlabeled)
2. Assign the b nearest neighbors to x_i to category i [if they have not already been assigned a label]
3. While Unlabeled is not empty:
 1. Choose x from Unlabeled randomly
 1. For $i=1, \dots, N$, let $D_i := 1 / (\min_{x' \in \text{Labeled}_i} D(x, x'))^{0.25}$
 2. Let $\text{Prob}(i) = \exp(D_i / t) / \sum_j \exp(D_j / t)$ be a softmax distribution
 3. Choose a label i from $\text{Prob}()$ and assign it to x

Step 2: Convexify

1. For each label i , ordered by increasing size:
 - a. Let M_i be $\text{ConvexHull}(\text{Labeled}_i) - \text{Labeled}_i$
 - b. Let R_i be the $c \cdot |M_i|$ closest points in M_i to Labeled_i
 - c. Relabel every point in R_i to label i

For the results reported in the paper, we set $b=100$, $N=7$, and used Euclidean distance. Our set of points X was 1331 points in CIELab space generated by taking the unit cube, with each axis partitioned at steps of interval 0.1, and converting these aRGB points to CIELab. The parameter t came from $\{5, 1, 0.1, 0.01, 0.001, 0.0005\}$ and c from $\{0, 0.25, 0.5, 0.75, 1\}$. We ran 10 trials for each pair of the t and c parameters, for a total of 300 trials. Fig. 3 shows random samples generated by this algorithm for various settings of the t and c parameters.



Fig. 3

Sample partitions of a 50x50 two-dimensional grid according to our color partitioning algorithm, showing the role of the two parameters t and c .

4.2.2 Computational Model

We used a multi-layer feed-forward neural network, with three input neurons (corresponding to dimensions in CIELab color space), $N=7$ output neurons, and two hidden layers of 12 units with the Exponential Linear activation function.¹

¹ For more details, see the code and data available online: <https://github.com/shanest/color-learning>.

4.2.3 Results

As seen in Fig. 4, there exists a strong correlation between degree of convexity and accuracy, supporting the claim that more convex color systems are easier to learn.

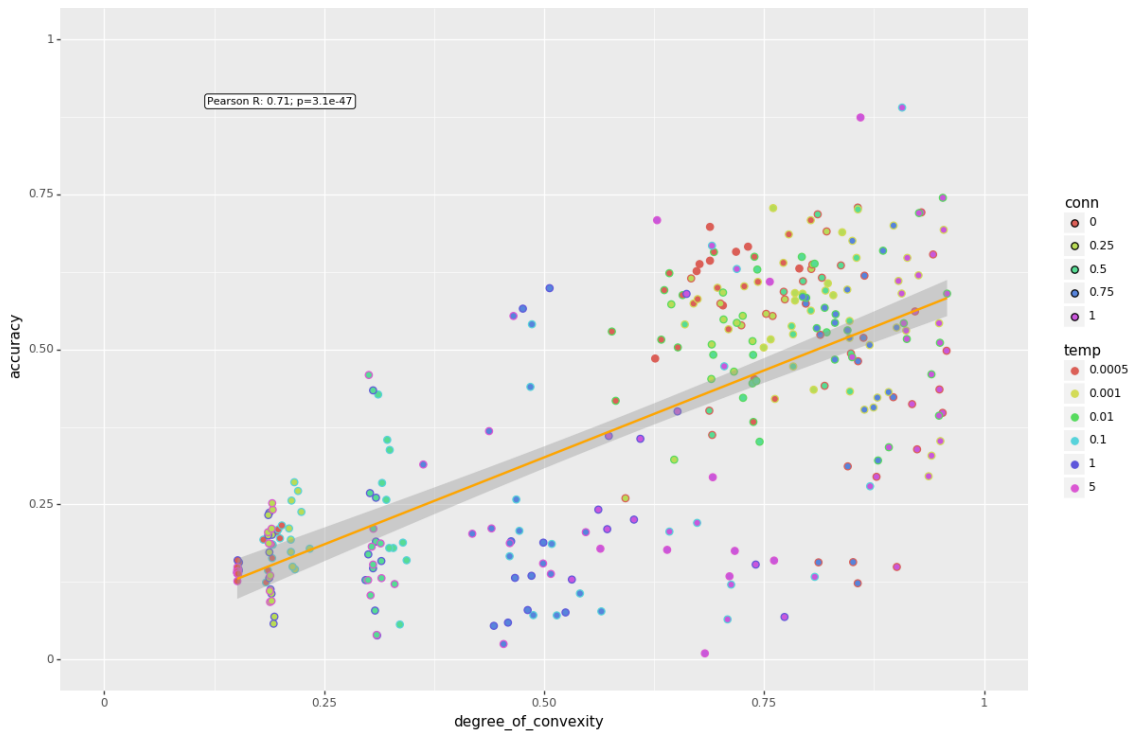


Fig. 4. Accuracy in classification by a multi-layer perceptron of artificial color systems, plotted against the degree of convexity of each such system. The dots are colored according to the c parameter; the border corresponds to the t parameter. These variables appear not to correlate with accuracy; detailed tests reported below confirm this.

In order to show that degree of convexity is the only variable that correlates with accuracy we ran a linear regression of accuracy against degree of convexity. We found Pearson's R to be 0.71 with $p=3.1e-47$. While the main effect was a correlation with degree of convexity, we also measured the effect of other variables. We included the parameters t and c from the algorithm, as well as their interaction $t*c$, and a few geometric properties of the color spaces: the size of the largest

(max) color, of the smallest (min) color, their difference (max-min) and ratio (max/min), as well as the median color size.

We did two things: first, we did a linear regression of accuracy on each of those variables independently. The results are reported in the first row of Table 1: degree of convexity was the only one with a significantly high value for Pearson's R^2 .

Next, we ran a multiple regression using *all* of the aforementioned variables. Then, for each variable, we ran a multiple regression with every variable *except for it*. The bottom row reports the difference in R^2 between those two regressions: the amount of variance explained by each variable when it is added last to a multiple regression. The second row of Table 1 shows that adding degree of convexity to all of the other variables explains a large amount of variance, while none of the others make a substantial independent contribution. These results show that degree of convexity does the main explanatory work in our main result.

	t	c	degree of convexity	max	min	max-min	max/min	median	$t*c$
R^2	0.193	0.002	0.504	0.004	0.018	0.001	0.001	0.006	0.064
Delta- R^2	0.0003	0.0561	0.4407	0.0	0.0	0.0	0.0024	0.0001	0.0013

Table 1. Statistics for color learning experiment: R^2 and Delta R^2 for several geometric variables.

5. Conclusions

It's natural to think that the languages of the world are shaped both by the cognitive makeup of language users and their local communicative needs, hence the nativism-relativism debate. Universals are more likely to be explained by cognitive makeup, since this does not vary in the way that communicative needs vary. We have argued that one aspect of our cognitive makeup -- the ease of learning -- can explain the presence of multiple semantic universals in disparate domains. This explanation makes minimal theory-laden assumptions and represents the first explanation of semantic universals that applies both to function words and to content words.

Acknowledgments

This work was funded by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

References:

1. L. M. Hyman, Universals in phonology. *Linguist. Rev.* **25**, 83–137 (2008).
2. N. Chomsky, *Aspects of the Theory of Syntax* (The MIT Press, Cambridge, 1965).
3. R. M. W. Dixon, Where have all the adjectives gone? *Stud. Lang.* **1**, 19–80 (1977).
4. W. Croft, *Typology and Universals* (Cambridge University Press, Cambridge, 1990).
5. J. Barwise, R. Cooper, Generalized Quantifiers and Natural Language. *Linguist. Philos.* **4**, 159–219 (1981).
6. C. Goddard, A. Wierzbicka, Eds., *Semantic and Lexical Universals* (John Benjamins Publishing Company, Amsterdam, 1994), vol. 25 of *Studies in Language Companion Series*.
7. C. Kemp, T. Regier, Kinship categories across languages reflect general communicative principles. *Science (80-.)*. **336**, 1049–1054 (2012).
8. C. Kemp, Y. Xu, T. Regier, Semantic typology and efficient communication. *Annu. Rev. Linguist.*, 1–23 (2018).
9. J. Feldman, Minimization of Boolean Complexity in Human Concept Learning. *Nature*. **407**, 630–633 (2000).
10. S. T. Piantadosi, J. B. Tenenbaum, N. D. Goodman, The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychol. Rev.* **123**, 392–424 (2016).
11. Y. Lecun, Y. Bengio, G. Hinton, Deep learning. *Nature*. **521**, 436–444 (2015).
12. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (The MIT Press, 2016; <https://www.deeplearningbook.org/>).
13. E. Sapir, The Status of Linguistics as a Science. *Language (Baltim)*. **5**, 207–214 (1929).
14. B. L. Whorf, Science and linguistics. *MIT Technol. Rev.* **42**, 229–231 (1940).

15. N. Evans, S. C. Levinson, The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–448 (2009).
16. S. Dehaene, *The Number Sense: How the Mind Creates Mathematics* (Oxford University Press, Oxford, 1997).
17. B. Berlin, P. Kay, *Basic Color Terms: Their Universality and Evolution* (University of California Press, 1969).
18. S. Peters, D. Westerstrahl, *Quantifiers in Language and Logic* (Clarendon Press, Oxford, 2006).
19. E. L. Keenan, J. Stavi, A Semantic Characterization of Natural Language Determiners. *Linguist. Philos.* **9**, 253–326 (1986).
20. J. van Benthem, Questions About Quantifiers. *J. Symb. Log.* **49**, 443–466 (1984).
21. P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, R. Cook, *World Color Survey* (CSLI Publications, 2010).
22. V. Loreto, A. Mukherjee, F. Tria, On the origin of the hierarchy of color names. *Proc. Natl. Acad. Sci.* **109**, 6819–6824 (2012).
23. T. Regier, P. Kay, Language, thought, and color: Whorf was half right. *Trends Cogn. Sci.* **13**, 439–446 (2009).
24. P. Gärdenfors, *Conceptual Spaces: the geometry of thought* (The MIT Press, Cambridge, MA, 2000).
25. P. Gärdenfors, *The Geometry of Meaning* (The MIT Press, 2014).
26. G. Jäger, in *Logic, Language, and Meaning: Amsterdam Colloquium 2009*, M. Aloni, H. Bastiaanse, T. de Jager, K. Schulz, Eds. (2010), pp. 11–20.
27. G. Jäger, The evolution of convex categories. *Linguist. Philos.* **30**, 551–564 (2007).