Frequency-Dependent Regularization in Syntactic Constructions

Zoey Liu
Boston College, ying.liu.5@bc.edu

Emily Morgan
University of California, Davis, eimorgan@ucdavis.edu

Follow this and additional works at: https://scholarworks.umass.edu/scil

Part of the Computational Linguistics Commons

Recommended Citation
DOI: https://doi.org/10.7275/6zr6-yt46
Available at: https://scholarworks.umass.edu/scil/vol4/iss1/42

This Abstract is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Frequency-Dependent Regularization in Syntactic Constructions

Zoey Liu  
Department of Computer Science  
Boston College  
ying.liu.5@bc.edu

Emily Morgan  
Department of Linguistics  
University of California, Davis  
eimorgan@ucdavis.edu

Previous research has suggested language users tend to have consistent ordering preferences given a syntactic construction with different grammatical alternatives (e.g. within the same binomial type, *theory and data* is consistently more preferred than *data and theory*). Borrowing the term of a well-known phenomenon in statistical learning, the tendency to make language structures more consistent is known as regularization. From the perspective of language learning and production, regularization means when there is variation in the input, language users will preferentially minimize the amount of variation in the output by reproducing the most frequent structure among all possible alternatives. For example, if a speaker has encountered *theory and data* 64 times and *data and theory* 36 times (a 1.8:1 ratio), regularization means the speaker will produce the former even more frequently than it was heard and thereby increases its output probability.

However, the tendency to regularize contradicts the dominant view from rational language comprehension (Levy, 2008), that since language users are sensitive to the probabilistic distributions of different structures, they will probability match rather than regularize. This means with a given construction as the input, language users will maintain the amount of variation of different alternatives in the output. For instance, with the same binomial type above, in production languages users will approximately maintain the 1.8:1 ratio (*theory and data* vs. *data and theory*) for the two structures as well.

Most experiments that tried to tease apart regularization and probability matching have focused on the learning and production of morphemes (Saldana et al., 2017), words (e.g. combining nouns with determiners) (Hudson Kam and Newport, 2005; Ferdinand et al., 2019; Smith et al., 2017), or word orders (Culbertson and Smolensky, 2012; Hudson Kam, 2019) only with artificially constructed stimuli. A few studies examined natural online production in sign languages, yet focusing mainly on morphological variation (e.g. Nicaraguan Sign Language (Senghas, 1995); American Sign language (Ross and Newport, 1996; Singleton and Newport, 2004). By contrast, explorations of regularization in syntactic constructions using naturalistic data have been lacking. Thus in general, whether language users tend to perform probability matching or regularization when reproducing structural variants and under what contextual conditions remain understudied.

Two notable exceptions thus far have attempted to narrow this gap, both using corpus data. Morgan and Levy (2016) demonstrated that the extent of regularization in binomial expressions in English is affected by the frequency of a binomial type: the higher the frequency, the stronger and more extreme preference there is for one alternative over the other (e.g. *safe and sound* >> *sound and safe*; *facts and techniques* > *techniques and facts*). In other words, regularization is frequency-dependent. Liu and Morgan (2020) further demonstrated this regularization bias in the dative construction in English (Bresnan et al., 2007), showing that the preference extremity within a dative type depends on its type frequency. Taking a data-driven approach, this study builds upon previous work and addresses two questions: (1) how wide-spread does frequency-dependent regularization exist in different syntactic construction types? (2) if this regularization bias were to hold regardless of the particular syntactic constructions, then what are the explanatory motivations behind it?

**Overall type frequency** We used as test cases adjective orders in phrases of the form Adj-Adj-N (AAN) and the dative constructions in English, the latter considered as a higher-level construction type than the former and the binomial types previously tested. An AAN type consists of two alternatives where the order of the two adjectives is
Figure 1: Selected results for AAN and datives with Levin class verbs: (a) demonstration of existence of frequency-dependent regularization (AAN: $\beta = 0.02$ (0.01, 0.02); dative: $\beta = 0.04$ (0.01, 0.07)); (b) simulating AAN corpus data using ILMs with different $R$ (a free parameter in the frequency-independent regularization function). Frequency-dependent regularization emerges with $R > 1$. All significance testing was from Bayesian linear regression predicting preference extremity as a function of overall type frequency.

![Figure 1](image.png)

Preference extremity was approximated as the probability of the more preferred structure given a construction type. Note the matter of interest here is how each construction type (e.g. verb idiosyncrasy for the dative constructions) affects regularization. Since the ordering preference of a dative construction is governed by different abstract constraints besides the verb (Bresnan et al., 2007), the effects of these constraints need to be excluded to more precisely quantify the role of the verb. To do that, we used mixed-effect models to predict the order, including phrasal length, definiteness, pronominality as fixed effects, and the verb as a random effect (models were fit to each of the three subsets described above). The probability for each structure within a dative type was computed based on just the random effect of the verb. Preference extremity was measured similarly for AAN; the model included the information-theoretic constraints (Futrell et al., 2020) as fixed effects and each AAN type as a random effect.

Iterated learning To address the evolution of frequency-dependent regularization, we used iterated learning models (ILM), which computationally simulate the process of how language structures evolve through generations. The learning process of ILM is a process of Bayesian inference. We augmented standard ILM in the same way as Liu and Morgan (2020), where a frequency-independent regularization function was applied in the data generation stage.

Results Results (e.g. Figure 1) show that frequency-dependent regularization exists in both AAN and the dative constructions with all three selections of verbs, and appropriate combinations of parameter settings in ILM give rise to this regularization bias. This indicates that language users regularize to some extent in production, and that interactions between language production and the continuous process of cultural transmission could lead to frequency-dependent regularization.
References


Kenny Smith, Amy Perfors, Olga Fehér, Anna Samara, Kate Swoboda, and Elizabeth Wonnacott. 2017. Language learning, language use and the evolution of linguistic variation. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1711):20160051.