Generalization and Probability Matching

Charles Yang
charles.yang@ling.upenn.edu

December 16, 2015

Probability matching is one of the most frequently observed behavior throughout the animal kingdom (Estes 1950, Bush and Mosteller 1951, Dawkins and Dawkins 1973, Herrnstein and Loveland 1975), and it is of little surprise that human language learners — children and adults, in both naturalistic and laboratory settings — can approximate the statistical distribution of linguistic variables (Roberts and Labov 1995, Roberts 1997, Kam and Newport 2005, Hudson-Kam and Newport 2009, Smith et al. 2009, Miller and Schmitt 2012). But probability matching in language appears to have some important and unique characteristics. A typical study of probability matching in animal behavior concerns the response to the stochastic nature of environmental stimuli, which tend to be *atomic*: turning left or right at the end of a T-maze, for instance. The probabilistic distribution of linguistic variables, by contrast, is almost always defined over a *class* of linguistic units such as words. For instance, the linguistic variable of English word-final t/d deletion (Labov and Cohen 1967) operates over a structurally specified set of words. Their phonological and grammatical environments have strong influence on the rate of deletion (Guy 1980), which in turn reflects general linguistic principles (Goldsmith 1976, Guy and Boberg 1997). In artificial language studies, learners also generalize probabilistic rules to novel items and constructions, although there are interesting differences between children and adult subjects (Kam and Newport 2005, Hudson-Kam and Newport 2009) which we address below.

The acquisition of probability matching in language, then, requires the learner to identify the structural conditioning of the linguistic variable as a *categorical* rule. That is, a rule is defined over a class of types, which are themselves subject to probabilistic application of the rule, as observed in the variant forms of a same type.

It is not obvious how language learners acquire probabilistic rules. Consider a hypothetical example, where a variable rule is defined over a set of $N = 1000$ items and yields two forms, A and B, with probability of 0.7 and 0.3 respectively. If the rule were to be used infinitely many times, each of the 1000 items will appear in form A 70% of time and form B 30% of time. But of course no one talks forever. Given the statistical distribution of linguistic items (Zipf 1949), the data generated by the variable rule must consist of:

(1) a. a very small fraction of the N items that appear in both form A and B, more or less reflecting the underlying probabilistic distribution of 70% and 30%,

b. a significant subset of the items which appear in both form A and B but have usage probabilities quite different from the "true" distribution,

c. a large subset of items that appear exclusively in form A or B, including many that are used only once and therefore cannot show variation.

A numerical simulation confirms these expectations. I generated $N = 1000$ "words" whose probabilities are defined by Zipf's Law. I then drew a sample of $S = 10{,}000$ tokens (with replacement): each time, the selected word surfaces in form A with probability 0.7 and in form B with probability 0.3. We observe:

(2) On average:
   a. about 900 out of the 1000 items are sampled at all;
   b. approximately 60% of those in (2a) sampled appear in both A and B;
   c. only about 30 out of those in (2b) show statistical variation that closely mirrors the true distribution (the percentage of A forms in the range of $[0.68, 0.72]$).

If the statistical distribution described in (1) and (2) is representative of naturalistic data, then the acquisition of probabilistic rules requires a very significant step of generalization. The learner must conclude that the behavior shown for a tiny fraction of the lexical items is to be extended to the entire set, including even those unattested in the learning input.

Similar situation arises in the acquisition of categorical rules which also appear to be "underrepresented" in the data. In earlier work, I investigated the status of syntactic rules in early child English (Yang 2013). A fully productive rule "*NP*→ *D N*" suggests that the determiner *D* (*the* and *a/n*) can be interchangeably used with singular nouns (*N*). In numerous corpus analyses, both children and adults produce fairly low values of combinatorial diversity: typically only 20-40% of nouns are paired with both determiners (Pine and Lieven 1997, Valian et al. 2009). Yet a rigorous statistical test shows that even very young children produce the level of diversity which, while low, is expected under a categorical rule that independently combines determiners and nouns (Yang 2013).

Consider a concrete case. Adam, an American English learning child studied by Roger Brown (1973), produced 3,729 determiner-noun combinations in his speech with 780 distinct nouns. Of these only 32.2% appeared with both determiners, which is similar to the expected value of 33.7% under the abstract *NP* rule. Adam's mother, who was recorded in the same corpus, produced a diversity measure of 30.3% out of 914 nouns. Even among the 469 nouns used at least twice, which provided opportunities to be used with both determiners, only over half (260) did so. A giant leap of faith is needed: Adam must — and apparently did at a very young age — generalize from a small subset of nouns with attested interchangeable determiners to all nouns.

A simple and promising solution comes from notion of "less is more" (Newport 1990, Elman 1993). Suppose that the child learner, due to maturational constraints and other factors, is only capable of attending to a subset of the input items. Specifically, let's assume that the child only retains and learns from the most frequent items, in effect ignoring those that are not sufficiently frequent. If so, the odds of acquiring a general rule improve considerably. First, more frequent items are more resistent to sampling effects and are thus more likely to reflect the full range of variation. Second, the statistical evidence for a rule is stronger when evaluated over a smaller number of types. In the speech transcript of Adam's mother, 43 of the top 50, 83 of the top 100, and 124 of the top 150 most frequent nouns are used with both determiners. These are very high batting averages for the *NP* rule: generalizing from 43 to 50 is more reasonable under any account of learning than generalizing 260 to 469 (or 914). Indeed, generalization over these high frequency items is valid according to the Tolerance Principle (Yang 2016), a mathematical model of inference. Once a categorical rule is acquired, the learner can generalize it to any lexical item, including those

that they have never encountered before. Importantly, the Tolerance Principle would not sanction a generalization if the all input items are included in the calculation (e.g., from 260 to 469).

The acquisition of probabilistic matching in language may follow a similar process. A young learner does not, perhaps can not, make use of all lexical items in the input: only a small subset of high frequency items such as those described in (1) and (2) constitutes the effective basis for generalization. Much like the case of determiner acquisition, the learner will observe that among this relevant subset, a great majority of items appear in both form *A* and *B*, thereby establishing the conclusion that the variable rule *R* categorically applies. Once this generalization is made, the learner can track the total frequencies of *A* and *B* *irrespective* of the lexical item. According to this view, a probabilistic rule is acquired first as a categorical one over a lexical class, before the probabilistic distribution of variation is established.[1]

Finally, the approach suggested here may account for some interesting differences between children and adults in the way that they form linguistic generalizations (Kam and Newport 2005, Hudson-Kam and Newport 2009, Schuler et al. 2015). For instance, Hudson-Kam and Newport (2009: Table 1) find that when presented with inconsistent rules of grammar, children, more so than adults, tend to overlook the lower frequency variants and become systematic users of the most dominant pattern. A conceivable account for this difference is that compared to adults, children only make use of a smaller vocabulary from the training stimuli. As noted earlier, a categorical rule is more likely to emerge when the denominator of generalization is smaller.

References

Brown, R. (1973). *A first language: The early stages*. Harvard University Press, Cambridge.

Bush, R. R. and Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 68(3):313–323.

Dawkins, R. and Dawkins, M. (1973). Decisions and the uncertainty of behaviour. *Behaviour*, 45(1/2):83–103.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological review*, 57(2):94.

Goldsmith, J. (1976). *Autosegmental phonology*. PhD thesis, MIT.

Guy, G. R. (1980). Variation in the group and the individual: The case of final stop deletion. In Labov, W., editor, *Locating language in time and space*, pages 1–35. Academic Press, New York.

Guy, G. R. and Boberg, C. (1997). Inherent variability and the Obligatory Contour Principle. *Language Variation and Change*, 9(2):149–164.

Herrnstein, R. J. and Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior*, 24:107–116.

[1]Such a learning model is not incompatible with lexically specific effects such as frequency: certain items may exceptionally deviate from the target probabilistic distribution. See Walker 2012, Labov 2014 for discussion of several well known sociolinguistic variables.

Hudson-Kam, C. L. and Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, 59(1):30–66.

Kam, C. H. and Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195.

Labov, W. (2014). The regularity of regular sound change. Manuscript in submission. University of Pennsylvania.

Labov, W. and Cohen, P. (1967). Systematic relations of standard and non-standard rules in the grammars of Negro speakers. In *Project Literacy Reports No. 8*, pages 66–84. Cornell University, Ithaca.

Miller, K. L. and Schmitt, C. (2012). Variable input and the acquisition of plural morphology. *Language Acquisition*, 19(3):223–261.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28.

Pine, J. M. and Lieven, E. V. (1997). Slot and frame patterns and the development of the determiner category. *Applied psycholinguistics*, 18(02):123–138.

Roberts, J. (1997). Acquisition of variable rules: A study of (-t, d) deletion in preschool children. *Journal of Child Language*, 24(2):351–372.

Roberts, J. and Labov, W. (1995). Learning to talk Philadelphian: acquisition of short *a* by preschool children. *Language Variation and Change*, 7:101–112.

Schuler, K., Davis, J., Yang, C., and Newport, E. (2015). Testing the Tolerance Principle for rule productivity in an artificial grammar. In *The Cognitive Development Society conference*, Columbus, OH.

Smith, J., Durham, M., and Fortune, L. (2009). Universal and dialect-specific pathways of acquisition: Caregivers, children, and t/d deletion. *Language Variation and Change*, 21(1):69–95.

Valian, V., Solt, S., and Stewart, J. (2009). Abstract categories or limited-scope formulae? the case of children's determiners. *Journal of Child Language*, 36(4):743–778.

Walker, J. A. (2012). Form, function, and frequency in phonological variation. *Language Variation and Change*, 24(03):397–415.

Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327

Yang, C. (2016). *Price of productivity: How children learn and break rules of language.* MIT Press, Cambridge, MA. Ms., University of Pennsylvania.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley, Cambridge.