The Design of Child Language

Charles Yang
University of Pennsylvania

April 2023

To learn a language, children must identify the infinitely generating rules that reside in a finite sample of data. To understand how language is learned, it is instructive to examine the properties of the input and the nature of rules formed on the basis of the input.

**

In this age of Big Data, it is easy to forget how little data is needed to learn a language. According to one estimate (Swingley 2009), children hear about one million sentences per year, most of which are remarkably simple and short, averaging only five words each. But even more remarkable than the scarcity is the *sparsity* of linguistic structures in the data. Words follow a *very* uneven statistical distribution (Zipf 1949, Yang 2013b). For example, the most frequent 100 words in English-learning children's input take up a whopping 62% of all the data; the rest, and the vast majority, appear rather sporadically, falling on a flat long tail. It goes without saying that the space of grammar is even more sparsely populated: rare words make even rarer combinations. But of course, the absence of evidence is not evidence of absence: rare or even unattested expressions may still be part of the language. Chomsky's *Colorless green ideas sleep furiously* is a famous example of a semantically nonsensical, statistically improbable, and yet perfectly grammatical sentence. Even the shortest and simplest structures are similarly affected by sparsity. For example, the English noun *sun* is almost always used with the article *the*, referring to the ball of hot plasma above, but no speaker will have trouble understanding or using *a sun*, *some suns*, etc. when other planetary systems are under discussion. Learning a language is to learn how to fill in the gaps, so to speak, that are left behind by the sparse input data.

****

Many have expressed skepticism in Chomsky's idealization of a homogeneous speech community (Chomsky 1965): surely there are innumerable ways in which our languages differ. It is perhaps even more remarkable that the most resounding confirmation for Chomsky's position comes from William Labov, the pioneer for the quantitative study of language variation, use, and change. Children acquire a strikingly uniform grammar within a speech community: "The end result is a high degree of uniformity in both the categorical and variable aspects of language production, where individual variation is reduced below the level of linguistic significance" (Labov 2012, 265; see also Labov 1972).

Recent years have seen renewed interest in individual differences across child learners (Kidd et al. 2018). As is obvious to this father of two, of course children are different from each other. Some like animals, some enjoy sports, some prefer to be quiet, and all are raised in their individual

families and social surroundings. But now we have a paradox: How do we reconcile variability at the level of the individual with uniformity at the level of the community? This tension is best illustrated by an examination of children's vocabulary – the main focus of recent studies of individual variation (Frank et al. 2021) – and the rules they learn on the basis of these vocabularies. Thanks to longitudinal records of child language development (e.g., Demuth et al. 2006), we can compare children's words as well as their rules. It turns out that on average, fewer than 20% of the first 100 words are shared between any two children. The overlap merely rises to about 30% for the first 500 words, and barely creeps above 40% for the first 1,000, which is an the upper limit of a three-year old's vocabulary size. Yet these children's *grammars* are highly uniform even by this very early stage of acquisition. Syntactic structures such as word order and productive combinations are firmly established as early as two (Brown 1973). And all children have learned the basic rules fo word formation: they even make the same errors such as *sleeped* and *foots*! Learning rules, therefore, cannot be dependent on learning specific words, contrary to claims in the usage-based learning literature (e.g., Tomasello 2003) that emphasize the role of memorization: a uniform learning outcome would be difficult to achieve.

*

Linguists are fond of saying "all grammars leak" (Sapir 1928, 38-39). The sentiment is easily verified in the English past tense system. The rule "add -ed" does not apply to some 150 irregular verbs, an essentially arbitrary list of historical residue, but nevertheless extends productively to novel words: when *google* became a verb in the 1990s, the past tense form *googled* was immediately available. Indeed, while the infinity of language has received justifiable attention, one should not overlook the disorderly corners of language, and the challenges that pose for language learning.

Language is always understood to have a component of arbitrariness thanks to de Saussure's observation about words. Yet arbitrariness seeps into all levels of language. For example, Mandarin Chinese has a few dozen noun classifiers: apart of the default classifier 个, most have to be individually memorized despite some semantic tendencies (Li and Thompson 1981). It is not uncommon to find heated discussion among native speakers about classifier choice for specific nouns: Which one is right for 拖拉机 (tractor)? 台 (for machines), or 辆 (for vehicles), or 架 (for carriages), or "it depends"? In Swedish, a language that marks gender, "tiger" and "chair" belong to one class whereas "lion" and "desk", words that cannot be more semantically similar, belong to another (Josefsson 2010). You just have to memorize them by rote.

The combinatorial process of language is also vulnerable: sometimes we have all the pieces of words but just can't put them together. Everyone knows that the past tense of the irregular verb *stride* is *strode* as in *I strode down the street*. But no one is certain about the past participle form: *I have _ down the street*, where *stridden*, *strode*, and *strided* all sound repellent (Pinker 1999). Most English auxiliary verbs allow the contraction of *not*: *is+not=isn't*, *do+not=don't*, *can+not=can't*, *will+not=won't*, *should+not=shouldn't*, etc. But American English speakers cannot use *amn't* or *mayn't* (for *am* and *may*), even though both the semantics and phonology of these forms are perfectly natural. Note that English has about the simplest morphology of all inflectional languages: missing forms multiply greatly in languages with more complex word formation processes (Halle 1973, Baerman and Corbett 2010).

The almighty syntax is not spared either. In English, there are about 200 verbs that can embed another sentence, expressing a wide range of meanings such as perception, mental state, and manner of speaking (Levin 1993):

(1) a. John saw that Bill ate the pizza.
b. John reasoned that Bill ate the pizza.
c. John whispered that Bill ate the pizza.

For most speakers, only *see* allows question formation all the way from the deep position of the object:

(2) a. What did John see that Bill ate _ ?
b. *What did John reason that Bill ate _ ?
c. *What did John whisper that Bill ate _ ?

Indeed, only some 30 embedding verbs are broadly accepted by native speakers as allowing such long-distance movements, with a good deal of variation across individuals. Learning these cannot be simpler: we just need to hear them used, without rhyme or reason (Chomsky 1977).

****

These major features of child language acquisition collectively reflect a central theme: language is a negotiation between systematicity and arbitrariness (Yang 2022). The linguistic experience for every child is a largely arbitrary sample of the language, which is in turn a largely arbitrary product of history. Yet from this arbitrariness will systematic rules emerge. The sparsity of data requires children to generalize very aggressively to create an infinite range of expressions — except when they don't.

Knowing when to generalize and when not to is the deepest puzzle in child language acquisition, and indeed all studies of learning by animals and machines alike. In an earlier phase of my work, and operating in the then orthodox theory according to which all grammatical options are innately available, I introduced simple domain-general probabilistic learning mechanisms to the selection of grammars (Yang 2002). But the rugged landscape of language, briefly reviewed earlier, points to severe shortcomings of that approach. It is also clear that not all options in grammar are available to children from the get-go. On the one hand, many aspects of child language are acquired quite late (Chomsky 1969, Gleitman and Gleitman 1979). On the other, children clearly exhibit stages of development demarcated by qualitative changes in their language. A well-known case is the acquisition of English past tense. For up to three years, children do not make mistakes when using irregular past tense. Then, and quite suddenly, overregularization errors such as *go-goed* and *sleep-sleeped* begin to appear (Ervin and Miller 1963). Since these forms are not found in the adult input, they must be children's own creation, signaling the emergence of the "add -ed" rule. But this rule is specific to the English language and cannot plausibly be innate: even grammatical tense marking is unlikely to be universal as can be seen in Chinese, Thai, Maori, Yoruba, and other so-called analytic languages.

It has since become clear that children must *form* the grammar, rather than simply picking one out of some preexisting set. Moreover, the grammar thus formed needs to provide a "good enough" coverage of the input data. In fact, the grammar cannot be more than merely good enough, for otherwise the highly variable individual experiences cannot be effectively smoothed out. At the same time, the grammar's reach should be restrained: there is always a significant amount of idiosyncrasy that needs no more than memorization. So sometimes language acquisition must grind to a halt. This is quite unexpected under the standard thinking about machine learning and cognitive science, which strives to provide the best account of the data – in terms of simplicity,

probability, coverage, or a composite of all the above: there will always a best hypothesis. Language provides an inconvenient counterpoint. As the examples of gaps and holes illustrate, *not* producing forms such as *amn't*, *stridden*, and those in (2b-c) is the correct outcome: even the best may not be good enough.

**

The Tolerance Principle (TP; Yang 2005, 2016) is a proposal for how children discover the rules of language but only where they can be found. It builds on the intuition that generalization requires a sufficient amount of supporting evidence: If there are 10 examples and all but one (9/10) support a rule, generalization ought to ensue. But no one in their right mind would extend a rule on the basis of 2/10: the learner should just memorize the two supporting examples. The solution is strikingly simple (proof due to Sam Gutmann):

(3) The Tolerance Principle
A pattern generalizes if e, the number of the items that do not follow the pattern, does not exceed a threshold, $\theta_N = N/\ln N$ where N is the cardinality of the item set.

For example, the emergence of "add -*ed*" rule is only possible when there is a sufficiently high number of *ed*-taking (i.e., regular) verbs to overcome the exceptions, the verbs that do not take -*ed* (i.e., the irregulars). This can only happen when the child's vocabulary size reaches certain value, as the English irregular verbs are also among the most frequent ones and will dominate the child's early vocabulary to prevent the formation of the "add -ed" rule. Hence children show the developmental pattern of memorization first and generalization later. The rise and fall of rules depends on the calibration of exceptions in children's developing vocabulary.

The TP meets the design features of child language acquisition. Table 1 provides some sample values of N and the associate threshold values θ_N. Note that as N increases, θ_N decreases quite sharply as a proportion. It follows that rules defined over a smaller vocabulary are easier to form. Furthermore, children's early vocabulary contains only highly frequent items, which have more opportunities to be attested in the general rules of the language, effectively reducing the sparsity problem. The proportionality criterion for rules under the TP normalizes individual variation to a uniform grammar: as long as the number of exceptions falls on the same side of the threshold, the resulting rule is the same no matter what the specific words are. Finally, a rule under the TP only needs to be adequate: when the tolerance threshold is not met, the learner will simply memorize the input without forming generalizations. Gaps and holes are simply cases where even the dominant pattern has more exceptions than tolerable.

There is now a considerable body of evidence in support of the TP, ranging from the acquisition of phonology, morphology, and syntax to the study of language variation and change. Additional confirmation comes from artificial language learning studies where the number of rule-following items and exceptions can be carefully manipulated. In particular, the findings that infants follow the TP in implicit learning tasks (Emond and Shi 2021) suggest that the TP is likely a formal principle of learning and generalization, one that is not restricted to rule formation in language.

****

So what happens to Universal Grammar and the prominence of nativism that followed Chomsky's revolution in linguistics and cognitive science?

It seems clear that Universal Grammar is, and certainly must be, a lot simpler than previously

N	θ_N	% of N
10	4	40.0
20	6	30.0
50	12	24.0
100	21	21.0
200	37	18.5
500	80	16.0
1000	144	14.4

Table 1: The maximum number of exceptions for a productive rule over N items

thought (Chomsky 1995, Hauser et al. 2002). What's innate is more likely a capacity to acquire grammars rather than having the grammars all baked in. This is an important distinction. The capacity to understand linear functions is clearly, and likely uniquely, within human grasp, but that's not to say that a two-year-old baby already has the concepts of slope and intercept. Similarly, Universal Grammar can only include the truly unique capacity of our species: Merge, or the formation of compositional and hierarchical structures (Yang 2013a, Berwick and Chomsky 2016). It would not contain noun, verbs, adjectives, prepositional phrases, etc. but rather the ability to create these categories and structures from the linguistic data that children receive. A mechanism of learning, one which may to operate in and out of language, will need to do most of the heavy lifting for child language acquisition (Yang et al. 2017).

References

Baerman, M. and Corbett, G. G. (2010). Defectiveness: Typology and diachrony. In Baerman, M., Corbett, G. G., and Brown, D., editors, *Defective paradigms: Missing forms and what they tell us*, pages 19–34. Oxford University Press, Oxford.

Berwick, R. C. and Chomsky, N. (2016). *Why only us: Language and evolution*. MIT Press, Cambridge, MA.

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

Chomsky, C. (1969). *The acquisition of syntax in children from 5 to 10*. MIT Press, Cambridge, MA.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Chomsky, N. (1977). On wh-movement. In Culicover, P. W., Wasow, T., and Akmajian, A., editors, *Formal syntax*, pages 71–132. Academic Press, New York.

Chomsky, N. (1995). *The minimalist program*. MIT Press, Cambridge, MA.

Demuth, K., Culbertson, J., and Alter, J. (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language and Speech*, 49(2):137–173.

Emond, E. and Shi, R. (2021). Infants' rule generalization is governed by the Tolerance Principle. In Dionne, D. and Vidal Covas, L.-A., editors, *Proceedings of the 45nd annual Boston University Conference on Language Development*, pages 191–204.

Ervin, S. M. and Miller, W. R. (1963). Language development. In Stevenson, H., editor, *Child psychology: The sixty-second yearbook of the National Society for the Study of Education*, pages 108–143. University of Chicago Press.

Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press, Cambridge, MA.

Gleitman, H. and Gleitman, L. (1979). Language use and language judgment. In Fillmore, C. J., Kemler, D., and Wang, W. S.-Y., editors, *Individual differences in language ability and language behavior*, pages 103–126. Elsevier.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry*, 4(1):3–16.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Josefsson, G. (2010). "disagreeing" pronominal reference in swedish and the interplay between formal and semantic gender. *Lingua*, 120(9):2095–2120.

Kidd, E., Donnelly, S., and Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, 22(2):154–169.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.

Labov, W. (2012). What is to be learned. *Review of Cognitive Linguistics*, 10(2):265–293.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Li, C. N. and Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books, New York.

Sapir, E. (1928). *Language: An introduction to the study of speech*. Harcourt Brace, New York.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3617–3632.

Tomasello, M. (2003). *Constructing a language*. Harvard University Press, Cambridge, MA.

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford University Press, Oxford.

Yang, C. (2005). On productivity. *Linguistic Variation Yearbook*, 5(1):333–370.

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

Yang, C. (2013b). Who's afraid of George Kingsley Zipf? Or: Do children and chimps have language? *Significance*, 10(6):29–34.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break rules of language*. MIT Press, Cambridge, MA.

Yang, C. (2022). Systematicity and arbitrariness in language: Saussurean rhapsody. In Papafragou, A., Trueswell, J., and Gleitman, L., editors, *The Oxford hanbook of the mental lexicon*, pages 327–355. Oxford University Press.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., and Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews*, 81(Part B):103 – 119.

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